SIAM J. MATH. ANAL.
Vol. 34, No. 1, pp. 1–27

DYNAMIC CONTACT WITH NORMAL COMPLIANCE WEAR AND DISCONTINUOUS FRICTION COEFFICIENT*

KENNETH L. KUTTLER† AND MEIR SHILLOR‡

Abstract. We apply the recent theory of evolution inclusions for set-valued pseudomonotone maps, developed in Kuttler and Shillor [*Commun. Contemp. Math.*, 1 (1999), pp. 87–123] to the problem of dynamic frictional contact with normal compliance and wear. The friction coefficient is assumed to be slip rate dependent, and may be continuous, or discontinuous in the form of a graph with a vertical segment at the origin, representing the transition from the static to the dynamic value. The wear of the contacting surfaces is modeled by the Archard law. We prove the existence of a weak solution for the problem. We establish the uniqueness of the weak solution in the case when the friction coefficient is continuous. We also show that the problem with prescribed wear depends continuously on the wear.

Key words. dynamic frictional contact, set-valued inclusions, existence and uniqueness, discontinuous friction coefficient, normal compliance, wear

AMS subject classifications. 74M10, 74M15, 35R35, 35R05, 35R70

PII. S0036141001391184

1. Introduction.

We use the theory of set-valued pseudomonotone maps, which we have developed in [18], to establish the existence of a weak solution of a dynamic frictional contact problem with wear, when the friction coefficient depends discontinuously on the slip velocity. The problem describes frictional dynamic contact between a deformable body, assumed to be viscoelastic, and a moving foundation and the resulting wear of the contact surface. This paper is a continuation of our investigation in [18], where the contact problem has been considered, however, with continuous coefficient of friction and without wear. The new features in the model are the description of friction with a discontinuous coefficient and inclusion of the wear of the contacting surfaces. We investigate the case when the friction coefficient jumps from a static value, when the contacting surfaces stick together, to the lower dynamic value at the onset of relative motion between them. Such a behavior is often assumed in engineering applications. The contact between the body and a moving rigid foundation is modeled with the normal compliance condition, and friction is modeled with the pressure dependent condition. We use the Archard law to describe the evolution of the wear. The problem is formulated as an abstract inclusion in a Banach space to which the results of [18] apply.

Dynamic frictional contact problems have been considered recently in [5, 6, 9, 19, 21, 26, 31], while quasi-static problems can be found in [2, 4, 7, 27, 30] and references to therein. See also [29] and the papers therein. It is a common assumption in engineering literature that the friction coefficient depends on the slip speed. However, there are only few and very recent mathematical publications which consider dynamic contact with a friction coefficient which depends on the slip velocity of the contacting surface [2, 10, 18, 19]. The last reference deals with a discontinuous slip-dependent

*Received by the editors June 20, 2001; accepted for publication (in revised form) April 16, 2002; published electronically August 15, 2002.
http://www.siam.org/journals/sima/34-1/39118.html
†Department of Mathematics, Brigham Young University, Provo, UT 84602 (klkuttle@math.byu.edu).
‡Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309 (shillor@oakland.edu).

coefficient, but in that problem the contact was assumed to be maintained and there was no wear. A simple one-dimensional dynamic problem was analyzed in [10], where a criterion for the appearance of dynamic instabilities was discovered. Analysis and numerical simulations of thermoelastic frictional contact of a beam were performed in [17]. The quasi-static problem with slip or total slip (over the contact history) dependent coefficient of friction can be found in [2] and the dynamic thermoviscoelastic problem in [3]. Frictional contact problems with wear can be found in [6] and [31].

In section 2, we present preliminary material which includes the abstract existence theorem of [18] that underlies our results here. The classical model for the process, its abstract formulation, the assumptions on the problem data, and the statement of our main result, Theorem 3.2, are given in section 3. Section 4 is devoted to approximate problems, with a known wear function, whose unique solvability, stated in Theorem 4.1, follows from the existence theorem in section 2. A solution of the contact problem with known wear, when the friction coefficient μ is continuous, is obtained as a limit of these approximate solutions in section 5. Under a mild additional assumption on the problem data we show that the solution is unique. We investigate in section 6 the case of a discontinuous friction coefficient. It is found that many of the necessary estimates do not depend on the continuity of μ, and this fact is exploited in establishing the existence of a weak solution in the case when μ has a jump discontinuity at the origin, when slip motion is initiated. The result is stated in Theorem 3.2, in the case when the wear is known. Uniqueness remains an unsolved problem in this case. In section 7, we prove the continuous dependence of the solutions of the problem on the wear function w. The result is stated in Theorem 7.1, and it has some merit on its own. In section 8, we deal with the problem with wear, which is assumed to evolve according to a local version of the Archard law. We use the results up to this point to establish the existence of the weak solution to the problem with wear; however, the questions of uniqueness and stability of the solutions remain open.

2. Preliminaries. The existence results to be presented in this paper are based on our recent theorems [18] for differential inclusions of the form

$$\left(B\left(t\right)u\left(t\right)\right)' + Au \ni f\left(t\right),$$

where A is a set-valued pseudomonotone map. Here, the prime denotes the time derivative which is understood in the sense of distributions. Let $\mathcal{V}$ be a reflexive Banach space, over $\mathbb{C}$, and let $\mathcal{V}'$ denote the space of conjugate linear maps. We start with (see, e.g., [22]) the following definition.

DEFINITION 2.1. *A map $A : \mathcal{V} \to \mathcal{P}\left(\mathcal{V}'\right)$ is said to be pseudomonotone if*

1. *the set Au is nonempty, bounded, closed, and convex for all $u \in \mathcal{V}$;*
2. *if F is a finite-dimensional subspace of $\mathcal{V}$, $u \in F$, and if U is a weakly open set in $\mathcal{V}'$ such that $Au \subseteq U$, then there exists $\delta > 0$ such that if $v \in B_\delta(u) \cap F$, then $Av \subseteq U$;*
3. *if $u_i \to u$ weakly in $\mathcal{V}$ and $u_i^* \in Au_i$ is such that*

$$\limsup_{i\to\infty} \mathrm{Re}\langle u_i^*, u_i - u\rangle_{\mathcal{V}} \leq 0, \tag{2.1}$$

then, for each $v \in \mathcal{V}$, there exists $u^*\left(v\right) \in Au$ such that

$$\liminf_{i\to\infty} \mathrm{Re}\langle u_i^*, u_i - v\rangle_{\mathcal{V}} \geq \mathrm{Re}\langle u^*(v), u - v\rangle_{\mathcal{V}}. \tag{2.2}$$

Here $B_\delta(u)$ denotes the ball of radius δ centered at u.

Our theory of set-valued evolution equations was developed in general reflexive Banach spaces. Here we restrict ourselves to the spaces which we now describe. Let $\Omega \subset \mathbb{R}^N$ $(N = 2, 3)$ be a domain, occupied by the deformable body, with Lipschitz boundary Γ. The surface is divided into three mutually disjoint parts Γ_D, Γ_N, and Γ_C such that $\Gamma_C \neq \emptyset$ is the potential contact surface. Next, we choose the space W as follows: if the body is clamped over Γ_D, with $\operatorname{meas}\Gamma_D > 0$, then we set $W = \{\mathbf{u} \in (H^1(\Omega))^N : \mathbf{u} = 0 \text{ on } \Gamma_D\}$; if the body is not held fixed ($\operatorname{meas}\Gamma_D = 0$), then $W = (H^1(\Omega))^N$. Now we let $q, p \geq 2$, set $D = W \cap (C^\infty(\overline{\Omega}))^N$, and define

$$\widetilde{V}_p = \{\mathbf{u} \in W : \gamma\mathbf{u} \in (L^p(\Gamma_C))^N\}, \tag{2.3}$$

with norm $||\mathbf{u}||_{\widetilde{V}_p} = ||\mathbf{u}||_W + ||\gamma\mathbf{u}||_{(L^p(\Gamma_C))^N}$, where $\gamma : W \to (L^2(\Gamma_C))^N$ is the trace operator. $\widetilde{V}_p$ is a reflexive Banach space since it is isometric to a closed subspace of $W \times (L^p(\Gamma_C))^N$. We denote by V_p the closure of D in $\widetilde{V}_p$. Then V_p is a reflexive Banach space, and for $p < q$

$$V_p \supseteq V_q, \qquad V_q \text{ is dense in } V_p. \tag{2.4}$$

Since V_p is dense in $H = (L^2(\Omega))^N$, we identify H and H' and write $V_p \subseteq H = H' \subseteq V_p'$. Let

$$\mathcal{V}_p = \{\mathbf{u} \in L^2(0, T; V_p) : ||\mathbf{u}||_{\mathcal{V}_p} < \infty\}, \tag{2.5}$$

equipped with norm

$$||\mathbf{u}||_{\mathcal{V}_p} = ||\mathbf{u}||_{L^2(0,T;W)} + ||\gamma\mathbf{u}||_{L^p(0,T;(L^p(\Gamma_C))^N)}. \tag{2.6}$$

$\mathcal{V}_p$ is a reflexive Banach space since it is isometric to a closed subspace of $L^2(0, T; W) \times L^p(0, T; (L^p(\Gamma_C))^N)$, and $\mathcal{V}_q$ is dense in $\mathcal{V}_p$ when $p < q$. Note that $\mathcal{V}_p' \subseteq L^{p'}(0, T; V_p')$ and the inclusion map is continuous.

Next, we define the Banach space X as follows:

$$X = \{\mathbf{u} \in \mathcal{V}_p : \mathbf{u}' \in \mathcal{V}_p'\}, \qquad ||\mathbf{u}||_X = ||\mathbf{u}||_{\mathcal{V}_p} + ||\mathbf{u}'||_{\mathcal{V}_p'}. \tag{2.7}$$

We shall use the following two results.

THEOREM 2.2 (see [20]). *Let $p \geq 1$, $q > 1$, $W \subseteq U \subseteq Y$ with compact inclusion map $i : W \to U$ and continuous inclusion map $i : U \to Y$ and let*

$$S_R = \{\mathbf{u} \in L^p(0, T; W) : \mathbf{u}' \in L^q(0, T; Y),\ ||\mathbf{u}||_{L^p(0,T;W)} + ||\mathbf{u}'||_{L^q(0,T;Y)} < R\}.$$

Then S_R is precompact in $L^p(0, T; U)$.

THEOREM 2.3 (see [28]). *Let W, U, and Y be as above and let*

$$S_{RT} = \{\mathbf{u} : ||\mathbf{u}(t)||_W + ||\mathbf{u}'||_{L^q(0,T;Y)} \leq R, \quad t \in [0, T]\}$$

for $q > 1$. Then S_{RT} is precompact in $C(0, T; U)$.

We now describe the abstract setting we shall use. Let V and W be reflexive Banach spaces over $\mathbb{C}$ and let $I = [a, b]$. We denote $\mathcal{W}_I \equiv L^2(I; W)$ and then $\mathcal{W}_I' = L^2(I; W')$. Also, when $I = [0, T]$, we write $\mathcal{V}$ instead of $\mathcal{V}_I$.

We assume that the family of operators $B(t)$ satisfies $B(t) \in \mathcal{L}(W, W')$ and

$$\langle B(t)\, u,\, v\rangle = \overline{\langle B(t)\, v, u\rangle}, \tag{2.8}$$

$$\langle B(t)\, u,\, u\rangle \geq 0, \tag{2.9}$$

$$B(t) = B(0) + \int_0^t B'(s)\, ds. \tag{2.10}$$

The operator L, associated with B, is defined as

$$D(L) \equiv \{u \in \mathcal{V} : (i^*Bu)' \in \mathcal{V}'\}, \tag{2.11}$$

$$Lu \equiv (i^*Bu)' \quad \text{for } u \in D(L), \tag{2.12}$$

where i is the inclusion map of V into W. The following lemma results from the definitions.

LEMMA 2.4. *L is a closed operator.*

We define

$$X \equiv D(L), \qquad ||u||_X \equiv ||Lu||_{\mathcal{V}'} + ||u||_{\mathcal{V}}.$$

By Lemma 2.4, X is isometric to a closed subspace of a product of reflexive Banach spaces and thus X is also reflexive. Under these conditions the following theorem was proved in [18].

THEOREM 2.5 (see [18]). *Let $u, v \in X$; then the following hold.*

1. *$t \to \langle B(t)u(t), v(t)\rangle_{W',W}$ equals an absolutely continuous function a.e. t, denoted by $\langle Bu, v\rangle(\cdot)$.*
2. *$\mathrm{Re}\langle Lu(t), u(t)\rangle = \frac{1}{2}[\langle Bu, u\rangle'(t) + \langle B'(t)u(t), u(t)\rangle]$ for a.e. t.*
3. *$|\langle Bu, v\rangle(t)| \le C||u||_X ||v||_X$ for some $C > 0$ and for all $t \in [0, T]$.*
4. *$t \to B(t)u(t)$ equals a function in $C(0, T; W')$, a.e. t, denoted by $Bu(\cdot)$.*
5. *$\sup\{||Bu(t)||_{W'},\ t \in [0, T]\} \le C||u||_X$ for some $C > 0$.*

If $K : X \to X'$ is given by

$$\langle Ku, v\rangle_{X',X} \equiv \int_0^T \langle Lu(t), v(t)\rangle dt + \langle Bu,\ v\rangle(0),$$

then

6. *K is linear, continuous, and weakly continuous.*
7. *$\mathrm{Re}\langle Ku, u\rangle = \frac{1}{2}[\langle Bu, u\rangle(T) + \langle Bu, u\rangle(0)] + \frac{1}{2}\int_0^T \langle B'(t)u(t), u(t)\rangle dt$.*

The operator A in the theorem and below is assumed to satisfy

$$A : \mathcal{V} \to \mathcal{P}(\mathcal{V}') \text{ is bounded;} \tag{2.13}$$

$$\liminf_{||u||_{\mathcal{V}} \to \infty} \frac{\{2\mathrm{Re}\langle u^*, u\rangle + \langle B'u, u\rangle + \langle Bu, u\rangle(T) : u^* \in Au\}}{||u||_{\mathcal{V}}} = \infty \tag{2.14}$$

for $u \in X$; and

$$A + K : X \to \mathcal{P}(X') \text{ is pseudomonotone.} \tag{2.15}$$

The following abstract theorem is the basis for the results in this paper.

THEOREM 2.6 (see [18]). *Let the spaces $\mathcal{V}$ and $\mathcal{W}$ be as defined above and let the operators $A : \mathcal{V} \to \mathcal{P}(\mathcal{V}')$ and $B(t)$ satisfy (2.13)–(2.15) and (2.10)–(2.12), respectively. If $f \in \mathcal{V}'$ and $u_0 \in W$, then there exists a solution $u \in \mathcal{V}$ to the initial value problem*

$$(i^*Bu)' + Au \ni f \text{ in } \mathcal{V}', \qquad Bu(0) = Bu_0 \text{ in } W'.$$

Here, i is the inclusion map $i : V \to W$. The proof of the theorem can be found in [18].

3. The model. We describe the classical problem and the assumptions on the data, then we formulate it abstractly, and we state our main results in Theorems 3.2–3.3. We use the isothermal version of the problem in [6] (see also [21, 8]). We refer the reader there for a more detailed description of the model. We use the *normal compliance* contact condition (see, e.g., [6, 5, 15, 13, 21, 27]) to describe the contact, together with a condition for dry friction. Dynamic problems with this condition have been investigated in [15, 5, 9, 6]. We use the Archard law, as has been done in [6], to describe the wear of the contact surface (see also [27, 30, 31]).

A viscoelastic body occupies the reference configuration $\Omega \subset \mathbb{R}^N$, with boundary surface $\Gamma = \partial\Omega$, such that $\Gamma = \overline{\Gamma}_C \cup \overline{\Gamma}_D \cup \overline{\Gamma}_N$. It may come in contact with a deformable moving foundation on the part Γ_C. We set $\Omega_T = \Omega \times (0,T)$ for $0 < T$ and denote the displacements vector by $\mathbf{u} = (u_1, \dots, u_N)$ and the stress tensor by $\sigma = \sigma(\mathbf{u}, \mathbf{u}') = (\sigma_{ij})$, where here and below $i, j = 1, \dots, N$, and a comma separates the components of a vector or tensor from partial derivatives.

The equations of motion, in dimensionless form, are

$$\mathbf{u}'' - \operatorname{Div} \sigma(\mathbf{u}, \mathbf{u}') = \mathbf{f}_B \qquad \text{in } \ \Omega_T, \tag{3.1}$$

where $\mathbf{f}_B$ represents the volume force acting on the body. Initially,

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \qquad \mathbf{u}'(\mathbf{x}, 0) = \mathbf{v}_0(\mathbf{x}) \ \text{ in } \Omega, \tag{3.2}$$

where $\mathbf{u}_0$ and $\mathbf{v}_0$ are the prescribed displacement and velocity fields, respectively.

The body is held fixed on Γ_D (when $\operatorname{meas} \Gamma_D \neq 0$) and tractions $\mathbf{f}_N$ act on Γ_N, thus

$$\mathbf{u} = 0 \quad \text{on } \Gamma_D, \qquad \sigma\mathbf{n} = \mathbf{f}_N \quad \text{on } \Gamma_N, \tag{3.3}$$

where $\mathbf{n}$ is the unit outward normal to Ω on Γ.

Our interest lies in the process on the contact surface Γ_C. We denote the normal component of the displacements vector on Γ by $u_n = \ \mathbf{u} \cdot \mathbf{n}$, the tangential components by $\mathbf{u}_T = \mathbf{u} - (\mathbf{u} \cdot \mathbf{n})\mathbf{n}$, the normal component of the traction by $\sigma_n = \sigma_{ij} n_j n_i$, and the tangential tractions by $\sigma_{Ti} = \sigma_{ij} n_j - \sigma_n n_i$.

We model the contact between the body and the foundation by the normal compliance condition. Let $g = g(\mathbf{x})$ be a nonnegative function on Γ_C, representing the gap between the body's surface (in the reference configuration) and the foundation, measured along the normal $\mathbf{n}$. We denote by $w = w(\mathbf{x}, t)$ the *wear function* which measures the wear of Γ_C at position $\mathbf{x}$ and time t. It describes the change in the surface, in the (negative) direction of the normal, resulting from material removal because of friction. We assume that the contact pressure is given by

$$\sigma_n = -p(u_n - w - g), \tag{3.4}$$

where $p(\cdot)$ is a nonnegative monotone function which vanishes for negative argument values. Thus, the pressure on the contact surface depends on the interpenetration $u_n - w - g$, when positive. The choice $p(r) = (r)_+^{m_n}$ can be found in [13, 21].

We note that as the wear of the surface increases the normal displacement needed for contact increases, too. In the tangential direction we employ a dry friction condition that is compatible with (3.4) and which has a slip dependent and discontinuous friction coefficient. Let μ^* denote the *friction graph*,

$$\mu^*(r) = \begin{cases} [\mu_d, \mu_s] & \text{when } r = 0, \\ \mu_c(r) & \text{when } r > 0, \end{cases} \tag{3.5}$$

where $r = |\mathbf{u}'_T - \mathbf{v}_*|$ denotes the relative slip between the surface and the foundation. Here, $\mathbf{v}_*$ is the tangential velocity of the foundation, and generally it depends on the location on the surface, thus it is assumed to lie in $L^\infty(0,T;L^\infty\left(\mathbb{R}^N\right))$. If the contact surface is flat, a portion of a plane, we may choose $\mathbf{v}_*$ to be a function of time only, but when the contact surface is not flat, even if the velocity of the foundation is constant, the tangential velocity is not constant and depends on the position and on time.

In the slip state ($0 < r$) the coefficient μ is given by $\mu_c(r)$, and $\mu_d = \lim_{s\to 0}\mu_c(s)$ denotes the dynamic value at zero slip. In the absence of relative slip μ may have any value in the interval $[\mu_d, \mu_s]$. Thus, we do not insist that it has the static value μ_s, although it is likely when the body is in stick state for a while. We assume that $\mu_c(r)$, for $r \geq 0$, is a given positive Lipschitz function which satisfies the conditions below.

Next, we consider the friction condition. As is well known in applications, and explained well in [25, 32], when the contact pressure is low to moderate, the real contact area is a small fraction of the nominal contact area, and the frictional tangential traction is proportional to the contact pressure, given by μp. This is the usual Coulomb's condition which is often used both in engineering and mathematical publications. However, when the contact pressure is very high, such as in metal forming processes, the fraction of the real contact area approaches unity, and the frictional traction reaches saturation and the maximal frictional resistance becomes independent of the contact pressure. Thus, there is a transition from the Coulomb law to the so-called Tresca law; see, e.g., [32]. Such a transition is observed both in elastic and plastic materials. A simple way to model such behavior is to introduce the truncated contact pressure function

$$p_{_R} = \begin{cases} p & \text{if } \ p \leq R, \\ R & \text{if } \ R \leq p. \end{cases}$$

Here, $R = $ const. is the pressure at which the friction traction levels off. We could have used, instead, a more general, and less transparent, function F such that $F = \mu p$ for small p, and asymptotically $F \to \mu R$ as $p \to \infty$.

Then the *friction bound* is defined as $\mu p_{_R}$, and the friction law is

$$\mu(|\mathbf{u}'_T - \mathbf{v}_*|) \in \mu^*(|\mathbf{u}'_T - \mathbf{v}_*|) \qquad \text{a.e. on } \Gamma_C, \tag{3.6}$$

$$|\sigma_T| \leq \mu_s p_{_R}(u_n - w - g), \tag{3.7}$$

$$\sigma_T = -\frac{\mathbf{u}'_T - \mathbf{v}_*}{|\mathbf{u}'_T - \mathbf{v}_*|}\mu_c\left(|\mathbf{u}'_T - \mathbf{v}_*|\right) p_{_R}(u_n - w - g) \quad \text{ if } \ \mathbf{u}'_T - \mathbf{v}_* \neq 0. \tag{3.8}$$

Conditions (3.6)–(3.8) model friction as follows. The tangential part of the traction is bounded by $\mu_s p_{_R}$. Sliding commences when $|\sigma_T|$ reaches the bound $\mu_s p_{_R}$, and then the tangential force has a direction opposite to the relative tangential velocity. The actual value of μ is a selection out of the graph, (3.6).

The contact surface Γ_C is divided, at each time instant, into the *separation*, *slip,* and *stick* zones.

We assume that the wear of the surface is either a given function or else it is proportional to the friction force and to the sliding speed, as in the Archard law,

$$\frac{\partial w}{\partial t} = k_w \mu_c(|\mathbf{u}'_T - \mathbf{v}_*|) p_{_R}(u_n - w - g) s_c(|\mathbf{u}'_T - \mathbf{v}_*|). \tag{3.9}$$

Here, k_w is a positive material constant, very small in practice. The function s_c is a regularization of $|\cdot|$ on $\mathbb{R}^N$ which is uniformly bounded and such that $s_c(r) = 0$ for $r = 0$. Note that we used μ_c in (3.9) since s_c vanishes when there is no slip.

The new features in the model are the dependence, which often can be observed experimentally, of the friction coefficient on the magnitude of the slip, $|\mathbf{u}'_T - \mathbf{v}_*|$, with a jump from a static to a dynamic value at the onset of sliding, and the wear of the contact surface. The problem with Lipschitz μ and without wear was investigated in [18].

Finally, we assume that the material is viscoelastic with constitutive law

$$\sigma = \sigma(\mathbf{u}, \mathbf{u}') = A\mathbf{u} + C\mathbf{u}', \tag{3.10}$$

i.e., $\sigma_{ij} = A_{ijkl}u_{k,l} + C_{ijkl}u'_{k,l}$, where the elasticity tensor A has the components A_{ijkl} and the viscosity tensor C has the components C_{ijkl}.

The classical formulation of the problem of *dynamic frictional contact with normal compliance wear and discontinuous slip dependent friction coefficient* is as follows: Find $\{\mathbf{u}, w\}$ such that (3.1)–(3.10) hold.

We make the following assumptions on the problem data. The normal pressure function $p(\cdot)$ is increasing and satisfies

$$|p(r_1) - p(r_2)| \le K(1 + r_1^{p-2} + r_2^{p-2})|r_1 - r_2|, \tag{3.11}$$

and either

$$0 \le p\,(r) \le K \text{ and } p = 2; \quad p(r) = 0, \qquad r < 0, \tag{3.12}$$

or

$$\delta^2 r^{p-1} - K \le p(r) \le K(1 + r^{p-1}), \qquad r \ge 0; \quad p(r) = 0, \qquad r < 0, \tag{3.13}$$

where $p \ge 2$ is a fixed exponent here and everywhere below, and δ and K are positive constants. Also, p is the exponent and $p(\cdot)$ is the normal compliance function. The choice made in [21] and [13] of $p(r) = r_+^{m_n}$, where $1 < m_T \le m_n$, corresponds to $p - 1 = m_n$ and clearly (3.13) holds for suitable constants K and δ. The function s_c satisfies

$$s_c(r) \le s_c^*, \qquad |s_c(r_1) - s_c(r_2)| \le \delta_c^* |r_1 - r_2|. \tag{3.14}$$

We assume that the coefficient of friction is a graph composed of the vertical segment $[\mu_d,\, \mu_s]$ and the function μ_c is bounded, positive, and Lipschitz continuous,

$$|\mu_c\,(r_1) - \mu_c\,(r_2)| \le \ \mathrm{Lip}_\mu\,|r_1 - r_2|\,, \qquad ||\mu_c||_{L^\infty} \le c_\mu. \tag{3.15}$$

We assume that the elasticity and viscosity coefficients A and C lie in $L^\infty(\Omega)$ and satisfy the following symmetries for $B = A$ or C:

$$B_{ijkl} = B_{ijlk}, \qquad B_{jikl} = B_{ijkl}, \qquad B_{ijkl} = B_{klij}, \tag{3.16}$$

and

$$B_{ijkl}\zeta_{ij}\zeta_{kl} \ge \lambda\zeta_{rs}\zeta_{rs}, \tag{3.17}$$

for all symmetric matrices ζ, where $0 < \lambda$.

We now obtain a weak formulation of problem (3.1)–(3.10) since, generally, the friction law and the set inclusion (3.6) preclude the existence of classical solutions.

We begin by defining the viscosity and elasticity operators $M, A : V_p \to V_p'$ as

$$\langle M\mathbf{u}, \mathbf{v}\rangle = \int_\Omega C_{ijkl}\mathbf{u}_{k,l}\mathbf{v}_{i,j}\,dx, \tag{3.18}$$

$$\langle A\mathbf{u}, \mathbf{v}\rangle = \int_\Omega A_{ijkl}\mathbf{u}_{k,l}\mathbf{v}_{i,j}\,dx. \tag{3.19}$$

It follows from our assumptions and Korn's inequality [23] that both M and A satisfy

$$\langle B\mathbf{u}, \mathbf{u}\rangle \geq \delta^2||\mathbf{u}||_W^2 - \lambda_0|\mathbf{u}|_H^2, \qquad \langle\, B\mathbf{u}, \mathbf{u}\rangle \geq 0, \qquad \langle B\mathbf{u}, \mathbf{v}\rangle = \langle B\mathbf{v}, \mathbf{u}\rangle$$

for $B = M$ or A, for some $\delta > 0$, and for $\lambda_0 \geq 0$.

The *normal compliance* operator $(\mathbf{v}, w) \to P(\mathbf{u}, w)$, which maps $\mathcal{V}_q \times L^p(\Gamma_C)$ to $\mathcal{V}_q'$ (for each $q \geq p$), is given by

$$\langle P(\mathbf{u}, w), \mathbf{z}\rangle = \int_0^T \int_{\Gamma_C} p(u_n - w - g)z_n\,d\Gamma dt, \tag{3.20}$$

where $\mathbf{u}(t) = \mathbf{u}_0 + \int_0^t \mathbf{v}(s)ds$, for $\mathbf{u}_0 \in V_q$. Next, we define $\mathbf{f} \in \mathcal{W}'$ as

$$\langle \mathbf{f}, \mathbf{z}\rangle_{\mathcal{W}',\mathcal{W}} = \int_0^T \int_\Omega \mathbf{f}_B\mathbf{z}\,dxdt + \int_0^T \int_{\partial\Omega} \mathbf{f}_N\gamma\mathbf{z}\,d\Gamma dt \tag{3.21}$$

for all $\mathbf{z} \in \mathcal{W}$. Here $\mathbf{f}_B$ represents a body force in $L^2(0, T; H)$ and $\mathbf{f}_N$ is a traction force in $L^2(0, T; L^2(\partial\Omega)^N)$.

Let $\gamma_T^* : L^{p'}(0, T; L^{p'}(\Gamma_C)^N) \to \mathcal{V}_p'$ be defined as

$$\langle \gamma_T^*\xi, \mathbf{w}\rangle = \int_0^T \int_{\Gamma_C} \xi \cdot \mathbf{w}_T\,d\Gamma dt.$$

The abstract form of the problem for the displacement $\mathbf{u}$, the velocity $\mathbf{v}$, and the wear w, is the following.

Problem $\mathcal{P}$. Find $\mathbf{u}, \mathbf{v} \in \mathcal{V}_p$, $w \in L^p(0, T; L^p(\Gamma_C))$ such that

$$\mathbf{v}' + M\mathbf{v} + A\mathbf{u} + P(\mathbf{u}, w) + \gamma_T^*\xi = \mathbf{f} \quad \text{ in } \quad \mathcal{V}_p', \tag{3.22}$$

$$w' = k_w\mu_c\left(|\mathbf{v}_T - \mathbf{v}_*|\right)p_R\left(u_n - w - g\right)s_c\left(|\mathbf{v}_T - \mathbf{v}_*|\right), \tag{3.23}$$

$$w(0) = 0, \qquad \mathbf{v}(0) = \mathbf{v}_0 \in H, \tag{3.24}$$

$$\mathbf{u}(t) = \mathbf{u}_0 + \int_0^t \mathbf{v}(s)\,ds, \qquad \mathbf{u}_0 \in V_p, \tag{3.25}$$

the inclusion (3.6) holds and for all $\mathbf{w} \in \mathcal{V}_p$,

$$\langle \gamma_T^*\xi, \mathbf{w}\rangle \leq \int_0^T \int_{\Gamma_C} \mu_c p_R\left(|\mathbf{v}_T - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_T - \mathbf{v}_*|\right)\,d\Gamma dt, \tag{3.26}$$

where $\mu_c = \mu_c(|\mathbf{v}_T - \mathbf{v}_*|)$ and $p_R = p_R(u_n - w - g)$.

When the triplet $\{\mathbf{u}, \mathbf{v}, w\}$ solves the abstract problem (3.22)–(3.26), then $\mathbf{u}$ and w are a weak solution of (3.1)–(3.10).

The main results in this paper are presented according to whether the wear is a given function or is determined by the differential equation (3.23). To begin with, we consider the following basic result, proved in section 5, in the case of a given wear function. We note that it includes all the published versions of the problem, such as [14, 21] or [15].

THEOREM 3.1. *Let $p \geq 2$ and let $w \in L^p(0,T;L^p(\Gamma_C))$, $w' \in L^p(0,T;L^p(\Gamma_C))$, $w' \geq 0$, $\mathbf{u}_0 \in V_p$, $\mathbf{v}_0 \in H$, $\mathbf{f} \in \mathcal{V}_p'$ and assume $\mu^*(r) = \mu_c(r)$, where μ_c is bounded and Lipschitz. Then there exists $\xi \in L^{p'}(0,T;L^{p'}(\Gamma_C)^N)$ and $\mathbf{v} \in L^2(0,T;W)$ such that*

$$(u_n - w - g)_+ \in L^\infty(0,T;L^p(\Gamma_C)), \tag{3.27}$$

$$\mathbf{v}' + M\mathbf{v} + A\mathbf{u} + P(\mathbf{u}, w) + \gamma_T^*\xi = \mathbf{f} \ \text{ in } \ \mathcal{V}_p', \tag{3.28}$$

$$\mathbf{v}(0) = \mathbf{v}_0\ ,\ \mathbf{u}(t) = \mathbf{u}_0 + \int_0^t \mathbf{v}(s)\,ds, \tag{3.29}$$

and

$$\langle \gamma_T^*\xi, \mathbf{w}\rangle \leq \int_0^T \int_{\Gamma_C} \mu_c p_R\left(|\mathbf{v}_T - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_T - \mathbf{v}_*|\right) d\Gamma\, dt, \tag{3.30}$$

where $\mu_c = \mu_c(|\mathbf{v}_T - \mathbf{v}_*|)$ and $p_R = p_R(u_n - w - g)$.

If, in addition, $p \leq 4$, then the solution $\{\mathbf{u}, \mathbf{v}\}$ is unique.

Next, we consider the case of a set-valued friction coefficient and given wear function. The proof can be found in section 6.

THEOREM 3.2. *Let $p \geq 2$ and let $\mathbf{u}_0 \in V_p$, $\mathbf{v}_0 \in H$, $\mathbf{f} \in \mathcal{V}_p'$, and $w, w' \in L^p(0,T;L^p(\Gamma_C))$ with $w' \geq 0$. Then there exists a pair $\{\mathbf{v}, \xi\}$ such that*

$$\mathbf{v} \in L^2(0,T;W), \qquad (u_n - w - g)_+ \in L^\infty(0,T;L^p(\Gamma_C)), \tag{3.31}$$

$$\mathbf{v}' + M\mathbf{v} + A\mathbf{u} + P(\mathbf{u}, w) + \gamma_T^*\xi = \mathbf{f} \ \text{ in } \ \mathcal{V}_p', \tag{3.32}$$

$$\mathbf{v}(0) = \mathbf{v}_0, \qquad \mathbf{u}(t) = \mathbf{u}_0 + \int_0^t \mathbf{v}(s)\,ds, \tag{3.33}$$

where ξ satisfies the inequality

$$\langle \gamma_T^*\xi, \mathbf{w}\rangle \leq \int_0^T \int_{\Gamma_C} \mu_c p_R\left(|\mathbf{v}_T - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_T - \mathbf{v}_*|\right) d\Gamma dt, \tag{3.34}$$

and where $\mu_c = \mu_c(|\mathbf{v}_T - \mathbf{v}_*|)$ and $p_R = p_R(u_n - w - g)$, for an element $(|\mathbf{v}_T - \mathbf{v}_*|, \mu(|\mathbf{v}_T - \mathbf{v}_*|))$ from the graph μ^*, a.e., and for all $\mathbf{w} \in \mathcal{V}_p$.

We note that this theorem guarantees only the existence of a solution. Indeed, it seems unreasonable to expect uniqueness when we have a graph in the problem; however, the question remains open.

Finally, we consider the case where the wear is a solution of the differential equation of Archard's law and $\mu^* = \mu_c$. This leads to the following theorem whose proof is in section 8.

THEOREM 3.3. *Assume* (3.12) *and that* $\mu = \mu_c = \mu^*$ *and* s_c *are bounded and Lipschitz continuous. Let* $\mathbf{u}_0 \in V_2, \mathbf{v}_0 \in H$, *and* $\mathbf{f} \in \mathcal{V}_2'$. *Then there exists a unique solution* $\{\mathbf{u}, \mathbf{v}, w\}$ *of problem* (3.22)–(3.26), *and it satisfies*

$$\mathbf{v} \in L^2(0,T;V_2), \quad \mathbf{v}' \in L^2(0,T;V_2'), \quad w,\, w' \in L^\infty(0,T;L^\infty(\Gamma_C)).$$

If we wish to take into account the possible dependence of μ and $p(\cdot)$ on the position x on the contact surface, all we need to do is to assume that both functions are measurable in x, in addition to the other assumptions above. This increase in generality is mainly technical and does not change any of the arguments and conclusions that follow. Therefore, we have omitted an explicit reference to it in the models.

Existence of weak solutions for the problem with friction graph and a wear function that is an unknown of the problem remains an important unresolved problem.

4. Approximate problems with given wear. In this section we consider regularized approximate problems in which w is a given function satisfying

$$w \in L^p(0,T;L^p(\Gamma_C)), \quad w' \in L^p(0,T;L^p(\Gamma_C)), \quad w' \geq 0,$$

and $\mu = \mu_c = \mu^*$ is a given Lipschitz continuous function of $|\mathbf{v}_T - \mathbf{v}_*|$.

First, let $\mathbf{u}_{0\varepsilon}$ be a sequence in D satisfying $\lim_{\varepsilon\to 0} \mathbf{u}_{0\varepsilon} = \mathbf{u}_0$ in V_p. We assume that $q = p^2(p-1)^{-1}$, thus

$$\frac{p-1}{q} + \frac{1}{p} + \frac{1}{q} = 1.$$

Next, let the operator J be defined by

$$\langle J\mathbf{u}, \mathbf{v}\rangle = \int_{\Gamma_C} ||\gamma\mathbf{u}||^{q-2}\gamma\mathbf{u}\cdot\gamma\mathbf{v}\, d\Gamma. \tag{4.1}$$

We use J to regularize problem (3.22)–(3.26) and for each $\varepsilon > 0$ the approximate problem is the following.

Problem $\mathcal{P}(\varepsilon)$. Find $\mathbf{v}_\varepsilon \in \mathcal{V}_q$ such that

$$\mathbf{v}_\varepsilon' + M\mathbf{v}_\varepsilon + A\mathbf{u}_\varepsilon + \varepsilon J\mathbf{v}_\varepsilon + P(\mathbf{u}_\varepsilon, w) + Q(\mathbf{v}_\varepsilon, w) \ni \mathbf{f} \ \text{ in } \mathcal{V}_q', \tag{4.2}$$

$$\mathbf{v}_\varepsilon(0) = \mathbf{v}_0 \in H, \tag{4.3}$$

$$\mathbf{u}_\varepsilon(t) = \mathbf{u}_{0\varepsilon} + \int_0^t \mathbf{v}_\varepsilon(s)\, ds. \tag{4.4}$$

Here, by $\mathbf{v}^* \in Q(\mathbf{v}, w) \subseteq \mathcal{V}_p'$ we mean that there exists $\mathbf{z} \in L^\infty(0,T;L^\infty(\Gamma_C)^N)$ such that

$$\langle \mathbf{v}^*, \mathbf{w}\rangle = \int_0^T \int_{\Gamma_C} \mu\left(|\mathbf{v}_T - \mathbf{v}_*|\right) p_R\left(u_n - w - g\right) \mathbf{z}\cdot\mathbf{w}_T\, d\Gamma dt, \tag{4.5}$$

and $\mathbf{z}$ satisfies

$$\int_0^T \int_{\Gamma_C} \mathbf{z}\cdot\mathbf{w}_T d\Gamma dt \leq \int_0^T \int_{\Gamma_C} \left(|\mathbf{v}_T - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_T - \mathbf{v}_*|\right) d\Gamma dt, \tag{4.6}$$

for all $\mathbf{w} \in \mathcal{V}_p$.

Below we omit the subscript ε for the sake of simplicity. We have the following result for the approximate problems.

THEOREM 4.1. *Assume that $p(\cdot)$ satisfies* (3.13). *Then for each $\varepsilon > 0$ there exists a solution $\mathbf{v}_\varepsilon \in \mathcal{V}_q$ of $\mathcal{P}(\varepsilon)$.*

The proof of the theorem is accomplished in a number of steps. We begin with the following assertion which follows directly from the definitions.

LEMMA 4.2. *The operators J, Q, M, A, and $P(\cdot, w)$ are bounded maps from $\mathcal{V}_q$ or $\mathcal{V}_q \times L^p(0,T; L^p(\Gamma_C))$ into $\mathcal{V}_q'$ or $\mathcal{P}(\mathcal{V}_q')$.*

Next, we change the dependent variable and set $\mathbf{y}e^{\lambda t} = \mathbf{v}$. Then, in terms of $\mathbf{y}$, the problem $\mathcal{P}(\varepsilon)$ consists of finding $\mathbf{y} \in \mathcal{V}_q$ such that

$$\mathbf{y}' + \lambda\mathbf{y} + M\mathbf{y} + e^{-\lambda(\cdot)}A\mathbf{u} + \varepsilon e^{-\lambda(\cdot)}J(e^{\lambda(\cdot)}\mathbf{y})$$

$$+ e^{-\lambda(\cdot)}P(\mathbf{u}, w) + e^{-\lambda(\cdot)}Q(e^{\lambda(\cdot)}\mathbf{y}, w) \ni e^{-\lambda(\cdot)}\mathbf{f} \qquad \text{in } \mathcal{V}_q', \tag{4.7}$$

$$\mathbf{y}(0) = \mathbf{v}_0 \in H. \tag{4.8}$$

Let X be the space given in (2.7). The next lemma will be used to show the operator Q_λ given by

$$\mathbf{y} \to Q_\lambda(\mathbf{y}, w) \equiv e^{-\lambda(\cdot)}Q(e^{\lambda(\cdot)}\mathbf{y}, w) \tag{4.9}$$

is pseudomonotone.

LEMMA 4.3. *If $\mathbf{v}^k \rightharpoonup \mathbf{v}$ in X, then $\gamma\mathbf{v}^k \to \gamma\mathbf{v}$ in $L^p(0,T;(L^p(\Gamma_C))^N)$.*

Proof. Since $p \geq 2$, it is straightforward to verify that if $\mathbf{v} \in X$, then $\mathbf{v}' \in L^{q'}(0,T;V_q')$ and $\mathbf{v}(t_1) - \mathbf{v}(t_2) = \int_{t_2}^{t_1} \mathbf{v}'(s)\,ds$. Let $W \subseteq U$ be such that the injection $W \to U$ is compact and $\gamma : U \to (L^2(\Gamma_C))^N$ is continuous. Since $\mathcal{V}_q$ embeds continuously into $L^2(0,T;W)$, Theorem 2.2 implies that $\mathbf{v}^k \to \mathbf{v}$ in $L^2(0,T;U)$. It follows that $\gamma\mathbf{v}^k \to \gamma\mathbf{v}$ in $L^2(0,T;(L^2(\Gamma_C))^N)$. Now if the lemma is not true, then there exists a sequence $\{\mathbf{v}^k\} \subseteq X$ such that $\mathbf{v}^k \rightharpoonup \mathbf{v}$ in X but $||\gamma\mathbf{v}^k - \gamma\mathbf{v}||_{L^p(0,T;(L^p(\Gamma_C))^N)} \geq \eta > 0$ for some η. By taking a subsequence, we may assume $\widetilde{\gamma\mathbf{v}}^k(\mathbf{x},t) \to \widetilde{\gamma\mathbf{v}}(\mathbf{x},t)$ a.e. $(\mathbf{x},t) \in \Gamma_C \times (0,T)$, since $\gamma\mathbf{v}^k \to \gamma\mathbf{v}$ in $L^2(0,T;(L^2(\Gamma_c))^N$. Here, " $\sim$ " means a product measurable representative. Since $\widetilde{\gamma\mathbf{v}}^k$ is bounded in $(L^q((0,T)\times\Gamma_C))^N$, the Fatou lemma guarantees that $\widetilde{\gamma\mathbf{v}}$ is also bounded in $L^q((0,T)\times\Gamma_C)$. Thus, the sequence $\{|\widetilde{\gamma\mathbf{v}}^k - \widetilde{\gamma\mathbf{v}}|^p\}$ is uniformly integrable, so it follows from the Vitali convergence theorem that

$$\lim_{k\to\infty} \int_{(0,T)\times\Gamma_C} |\widetilde{\gamma\mathbf{v}}^k - \widetilde{\gamma\mathbf{v}}|^p \, d\Gamma dt = 0.$$

This contradicts the assumption that $||\gamma\mathbf{v}^k - \gamma\mathbf{v}||_{L^p(0,T;(L^p(\Gamma_C))^N)} \geq \eta > 0$ and thus proves the lemma.

LEMMA 4.4. *If $\mathbf{y}^k \rightharpoonup \mathbf{y}$ in X, then*

$$p(u_n^k - w - g) \to p(u_n - w - g) \quad \text{in } L^{p'}(0,T;L^{p'}(\Gamma_C)) \tag{4.10}$$

and

$$\mu\left(|\mathbf{v}_T^k - \mathbf{v}_*|\right) \to \mu\left(|\mathbf{v}_T - \mathbf{v}_*|\right) \quad \text{in } L^p(0,T;L^p(\Gamma_C)). \tag{4.11}$$

Proof. To simplify the notation we let $F = p(u_n - w - g)$, $F^k = p(u_n^k - w - g)$, $\mu = \mu(|\mathbf{v}_T - \mathbf{v}_*|)$, and $\mu^k = \mu(|\mathbf{v}_T^k - \mathbf{v}_*|)$. Now it follows from (3.13) that

$$\left|F^k - F\right| \leq K\left(1 + \left|u_n^k\right|^{p-2} + |u_n|^{p-2}\right)\left|u_n^k - u_n\right|.$$

We will show that $|u_n^k|^{p-2}|u_n^k-u_n| \to 0$ in $L^{p'}(0,T;L^{p'}(\Gamma_C))$ and observe that simpler arguments apply to the other two terms. We have

$$\begin{aligned}
&\int_0^T \int_{\Gamma_C} \left|u_n^k\right|^{(p-2)p'} \left|u_n^k - u_n\right|^{p'} \, d\Gamma dt \\
&\quad \leq \left(\int_0^T \int_{\Gamma_C} \left|u_n^k - u_n\right|^p d\Gamma dt \right)^{p'/p} \left(\int_0^T \int_{\Gamma_C} \left|u_n^k\right|^p \, d\Gamma dt \right)^{(p-p')/p} \\
&\quad \leq c \left(\int_0^T \int_{\Gamma_C} \left|u_n^k - u_n\right|^p d\Gamma dt \right)^{p'/p},
\end{aligned}$$

which converges to zero by Lemma 4.3. Moreover,

$$\int_0^T \int_{\Gamma_C} |\mu^k - \mu|^p \, d\Gamma dt \leq C \ \ \mathrm{Lip}_\mu^p \int_0^T \int_{\Gamma_C} \left|\gamma \mathbf{v}^k - \gamma \mathbf{v}\right|^p \, d\Gamma dt,$$

which also converges to zero by Lemma 4.3. The other terms behave similarly.

LEMMA 4.5. *Let* $\mathbf{y}^k \rightharpoonup \mathbf{y}$ *in* X *and* $\mathbf{z}^k \rightharpoonup \mathbf{z}$ *in* $L^\infty(0,T;L^\infty(\Gamma_C)^N)$. *If* $\mathbf{w} \in L^p(0,T;L^p(\Gamma_C)^N)$, *then*

$$\int_0^T \int_{\Gamma_C} F^k \mu^k \, \mathbf{z}^k \cdot \mathbf{w}_T \, d\Gamma dt \to \int_0^T \int_{\Gamma_C} F\mu \ \ \mathbf{z} \cdot \mathbf{w}_T \, d\Gamma dt. \tag{4.12}$$

Proof. We argue by contradiction. If (4.12) does not hold, then there exist two sequences $\mathbf{y}^k \rightharpoonup \mathbf{y}$ in X and $\mathbf{z}^k \rightharpoonup \mathbf{z}$ in $L^\infty(0,T;L^\infty(\Gamma_C)^N)$ and $\mathbf{w} \in L^p(0,T;L^p(\Gamma_C)^N)$ such that

$$\left| \int_0^T \int_{\Gamma_C} F^k \mu^k \, \mathbf{z}^k \cdot \mathbf{w}_T \, d\Gamma dt - \int_0^T \int_{\Gamma_C} F\mu \, \mathbf{z} \cdot \mathbf{w}_T \, d\Gamma dt \right| \geq 2\hat{\varepsilon}.$$

Since $L^\infty(0,T;L^\infty(\Gamma_C)^N)$ is dense in $L^p(0,T;L^p(\Gamma_C)^N)$, we find that, for $\mathbf{w} \in L^\infty(0,T;L^\infty(\Gamma_C)^N)$,

$$\left| \int_0^T \int_{\Gamma_C} F^k \mu^k \, \mathbf{z}^k \cdot \mathbf{w}_T \, d\Gamma dt - \int_0^T \int_{\Gamma_C} F\mu \, \mathbf{z} \cdot \mathbf{w}_T \, d\Gamma dt \right| \geq \hat{\varepsilon}. \tag{4.13}$$

However, by Lemma 4.4, $\mu(|\mathbf{v}_T^k - \mathbf{v}_*|)p(u_n^k - w - g) \to \mu(|\mathbf{v}_T - \mathbf{v}_*|)p(u_n - w - g)$ in $L^1(0,T;L^1(\Gamma_C))$. Therefore, (4.13) cannot hold for all k, which proves the lemma.

LEMMA 4.6. Q_λ *is a bounded pseudomonotone operator.*

Proof. We have already observed that Q_λ is bounded, and it is straightforward to show that $Q_\lambda(\mathbf{y})$ is convex. Suppose that $Q_\lambda(\mathbf{y}) \subseteq U$, where U is a weakly open set in X', that $\mathbf{y}_k^* \in Q_\lambda(\mathbf{y}) \setminus U$, and that $\mathbf{y}^k \rightharpoonup \mathbf{y}$ in X, where $\mathbf{y}_k^* \in Q_\lambda\left(\mathbf{y}^k\right)$. Let $U_\lambda \equiv e^{\lambda(\cdot)}U$; then U_λ is weakly open in X' containing $Q(\mathbf{v})$, $\mathbf{v}^k \rightharpoonup \mathbf{v}$ in X, and $\mathbf{v}_k^* \equiv e^{\lambda(\cdot)}\mathbf{y}_k^* \in Q(\mathbf{v}^k) \setminus U_\lambda$. Next, let $\{\mathbf{z}^k\}$ be a sequence in $L^\infty(0,T;L^\infty(\Gamma_C)^N)$ as in the definition of Q such that, possibly for a subsequence, $\mathbf{z}^k \rightharpoonup \mathbf{z}$ in $L^\infty(0,T;L^\infty(\Gamma_C)^N)$.

From Lemma 4.3,

$$\begin{aligned}\int_0^T \int_{\Gamma_C} \mathbf{z}\cdot\mathbf{w}_T\, d\Gamma dt &= \lim_{k\to\infty} \int_0^T \int_{\Gamma_C} \mathbf{z}^k\cdot\mathbf{w}_T\, d\Gamma dt \\ &\le \lim_{k\to\infty} \int_0^T \int_{\Gamma_C} \left(|\ \mathbf{v}_T^k - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_T^k - \mathbf{v}_*|\right)\, d\Gamma dt \\ &\le \int_0^T \int_{\Gamma_C} \left(|\mathbf{v}_T - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_T - \mathbf{v}_*|\right)\, d\Gamma\, dt.\end{aligned}$$

Now, using the notation $p_{_R}^k = p_{_R}(u_n^k - w - g)$ and $p_{_R} = p_{_R}(u_n - w - g)$, we have

$$\langle \mathbf{v}_k^*, \mathbf{w}\rangle \equiv \int_0^T \int_{\Gamma_C} \mu\left(\left|\mathbf{v}_T^k - \mathbf{v}_*\right|\right) p_{_R}^k\, \mathbf{z}^k\cdot\mathbf{w}_T\, d\Gamma dt,$$

and so from Lemma 4.5 we know that $\mathbf{v}_k^* \rightharpoonup \mathbf{v}^*$, where

$$\langle \mathbf{v}^*, \mathbf{w}\rangle \equiv \int_0^T \int_{\Gamma_C} \mu\left(|\mathbf{v}_T - \mathbf{v}_*|\right) p_{_R}\, \mathbf{z}\cdot\mathbf{w}_T\, d\Gamma dt. \tag{4.14}$$

Thus, $\mathbf{v}^* \in Q(\mathbf{v}) \subseteq U_\lambda$ by the definition of Q. This contradicts the assumption that $\mathbf{v}_k^* \notin U_\lambda$ for all k, and hence $Q(\mathbf{v}^k) \subseteq U_\lambda$ for all large k. This argument also shows that $Q_\lambda(\mathbf{y})$ is closed. It remains to verify conditions (2.1) and (2.2).

To that end let $\mathbf{y}^k \rightharpoonup \mathbf{y}$ and $\mathbf{y}_k^* \in Q_\lambda(\mathbf{y}^k)$. We show that if $\mathbf{w} \in X$, then

$$\liminf_{k\to\infty} \langle \mathbf{y}_k^*, \mathbf{y}^k - \mathbf{w}\rangle \ge \langle \mathbf{y}^*(\mathbf{w}), \mathbf{y} - \mathbf{w}\rangle, \qquad \mathbf{y}^*(\mathbf{w}) \in Q_\lambda(\mathbf{y}).$$

We choose a subsequence $\mathbf{y}^k$ (depending on $\mathbf{w}$) such that

$$\lim_{k\to\infty} \langle \mathbf{y}_k^*, \mathbf{y}^k - \mathbf{w}\rangle = \liminf_{k\to\infty} \langle \mathbf{y}_k^*, \mathbf{y}^k - \mathbf{w}\rangle.$$

For $\mathbf{v}_k^* = e^{\lambda(\cdot)}\mathbf{y}_k^* \in Q\left(\mathbf{v}^k\right)$ we let $\mathbf{z}^k \in L^\infty(0,T;L^\infty(\Gamma_C)^N)$ be as in the definition of Q. We take a further subsequence, if necessary, such that $\mathbf{z}^k \rightharpoonup \mathbf{z}$ in $L^\infty(0,T;L^\infty(\Gamma_C)^N)$. Then $\mathbf{z}$ satisfies (4.6) by Lemma 4.3. It follows from Lemma 4.5 that if we define $\mathbf{y}^*(\mathbf{w})$ by

$$\langle \mathbf{y}^*(\mathbf{w}), \mathbf{b}\rangle = \int_0^T \int_{\Gamma_C} e^{-\lambda t} p_{_R}\left(u_n - w - g\right) \mu\left(|\mathbf{v}_T - \mathbf{v}_*|\right) \mathbf{z}\cdot\mathbf{b}_T\, d\Gamma dt,$$

then

$$\begin{aligned}\liminf_{k\to\infty} \langle \mathbf{y}_k^*, \mathbf{y}^k - \mathbf{w}\rangle &= \lim_{k\ \to\infty} \langle \mathbf{y}_k^*, \mathbf{y}^k - \mathbf{w}\rangle \\ &= \lim_{k\to\infty} \int_0^T \int_{\Gamma_C} e^{-\lambda t}\mu^k\, p_{_R}^k\, \mathbf{z}^k\cdot\left(\mathbf{y}_T^k - \mathbf{w}_T\right) d\Gamma dt \\ &= \int_0^T \int_{\Gamma_C} e^{-\lambda t}\mu\, p_{_R}\, \mathbf{z}\cdot\left(\mathbf{y}_T - \mathbf{w}_T\right) d\Gamma dt = \langle \mathbf{y}^*(\mathbf{w}), \mathbf{y} - \mathbf{w}\rangle.\end{aligned}$$

This proves the lemma.

LEMMA 4.7. *If $\mathbf{v}^k \rightharpoonup \mathbf{v}$ in X, then $P(\mathbf{u}^k, w) \to P(\mathbf{u}, w)$ in $\mathcal{V}_q'$.*

Proof. Let $\mathbf{w} \in \mathcal{V}_q$. Then we have from the definition of P and (3.13) that

$$\begin{aligned}
&\left|\left\langle P(\mathbf{u}^k, w) - P(\mathbf{u}, w), \mathbf{w}\right\rangle\right| \\
&\quad \le K \int_0^T \int_{\Gamma_C} (1 + |u_n^k|^{p-2} + |u_n|^{p-2})|u_n^k - u_n||w_n| \, d\Gamma dt \\
&\quad \le K \int_0^T \left(\int_{\Gamma_C} (1 + |u_n^k|^p + |u_n|^p) \, d\Gamma\right)^{\frac{p-2}{p}} \left(\int_{\Gamma_C} |u_n^k - u_n|^p \, d\Gamma\right)^{\frac{1}{p}} \\
&\qquad\qquad \times \left(\int_{\Gamma_C} |w_n|^p d\Gamma\right)^{\frac{1}{p}} dt \\
&\quad \le K ||u_n^k - u_n||_{L^p(0,T;(L^p(\Gamma_C))^N)} || \, \mathbf{w}||_{\mathcal{V}_q}.
\end{aligned}$$

Thus, $||P(\mathbf{u}^k, w) - P(\mathbf{u}, w)||_{\mathcal{V}_q'} \le K||\gamma \mathbf{u}^k - \gamma \mathbf{u}||_{L^p(0,T;(L^p(\Gamma_C))^N)}$, and the result follows from Lemma 4.3.

Now for each $\lambda \ge 0$ the map $\mathbf{y} \to e^{-\lambda(\cdot)} A\mathbf{u}$ is monotone; in fact,

$$\begin{aligned}
\left\langle e^{-\lambda(\cdot)} A(\mathbf{u}_1 - \mathbf{u}_2), \mathbf{y}_1 - \mathbf{y}_2 \right\rangle &= \frac{1}{2} \int_0^T e^{-2\lambda t} \frac{d}{dt} \left\langle A(\mathbf{u}_1 - \mathbf{u}_2), \mathbf{u}_1 - \mathbf{u}_2 \right\rangle dt \\
&= \frac{1}{2} e^{-2\lambda T} \left\langle A(\mathbf{u}_1(T) - \mathbf{u}_2(T)), \mathbf{u}_1(T) - \mathbf{u}_2(T) \right\rangle \\
&\quad + \lambda \int_0^T \left\langle A(\mathbf{u}_1 - \mathbf{u}_2), \mathbf{u}_1 - \mathbf{u}_2 \right\rangle e^{-2\lambda t} dt.
\end{aligned} \tag{4.15}$$

Also, the map $\mathbf{y} \to \varepsilon e^{-\lambda(\cdot)} J(e^{\lambda(\cdot)} \mathbf{y})$ is monotone. Next, $\mathbf{y}^k \rightharpoonup \mathbf{y}$ in X if and only if $\mathbf{v}^k \rightharpoonup \mathbf{v}$ in X, and Lemma 4.7 implies that the operator $\mathbf{y} \to e^{-\lambda(\cdot)} P(\mathbf{u}, w)$ is completely continuous; and if we let

$$\begin{aligned}
\mathcal{A}_\lambda \mathbf{y} &= \lambda \mathbf{y} + M\mathbf{y} + e^{-\lambda(\cdot)} A(\mathbf{u}) + \varepsilon e^{-\lambda(\cdot)} J(e^{\lambda(\cdot)} \mathbf{y}) \\
&\quad + e^{-\lambda(\cdot)} Q(e^{\lambda(\cdot)} \mathbf{y}) + e^{-\lambda(\cdot)} P(\mathbf{u}, w),
\end{aligned} \tag{4.16}$$

then $\mathcal{A}_\lambda$ is a sum of bounded pseudomonotone operators. Consequently, $\mathcal{A}_\lambda : X \to \mathcal{P}(X')$ is pseudomonotone [22], verifying condition (2.15) for $\mathcal{A}_\lambda$. We now check the coercivity of $\mathcal{A}_\lambda$ (2.14). To this end, we consider the various terms of $\langle \mathcal{A}_\lambda \mathbf{y}, \mathbf{y} \rangle$. Let $\mathbf{y}^* \in Q_\lambda(\mathbf{y})$, which implies that $\mathbf{y}^* \in e^{-\lambda(\cdot)} Q(e^{\lambda(\cdot)} \mathbf{v})$ and so $\mathbf{y}^* = e^{-\lambda(\cdot)} \mathbf{v}^*$, where $\mathbf{v}^* \in Q\left(e^{\lambda(\cdot)} \mathbf{v}\right)$. Therefore,

$$\langle \mathbf{y}^*, \mathbf{y} \rangle = \langle e^{-\lambda(\cdot)} \mathbf{v}^*, e^{\lambda(\cdot)} \mathbf{v} \rangle = \langle \mathbf{v}^*, \mathbf{v} \rangle = \int_0^T \int_{\Gamma_C} \mu p_{_R} \, \mathbf{z} \cdot \mathbf{v}_T \, d\Gamma dt,$$

where $p_{_R} = p_{_R}(u_n - w - g)$ and $\mathbf{z} \in L^\infty(0,T; L^\infty(\Gamma_C)^N)$ satisfies

$$\int_0^T \int_{\Gamma_C} \mathbf{z} \cdot \mathbf{w}_T \, d\Gamma dt \le \int_0^T \int_{\Gamma_C} \left(|e^{\lambda(\cdot)} \mathbf{v}_T - \mathbf{v}_* + \mathbf{w}_T| - |e^{\lambda(\cdot)} \mathbf{v}_T - \mathbf{v}_*| \right) d\Gamma dt, \tag{4.17}$$

and $\mathbf{u}(t) = \mathbf{u}_0 + \int_0^t e^{\lambda s} \mathbf{v}(s) \, ds$. Thus,

$$\begin{aligned}
\langle \mathbf{y}^*, \mathbf{y} \rangle &= \int_0^T \int_{\Gamma_C} e^{-\lambda t} \mu p_{_R} \, \mathbf{z} \cdot \left(e^{\lambda t} \mathbf{v}_T - \mathbf{v}_* \right) d\Gamma dt \\
&\quad + \int_0^T \int_{\Gamma_C} e^{-\lambda t} \mu p_{_R} \, \mathbf{z} \cdot \mathbf{v}_* \, d\Gamma dt.
\end{aligned} \tag{4.18}$$

Now the first integral is nonnegative by a routine argument involving (4.17), and since $\mathbf{v}_* \in L^\infty(0,T;L^p(\Gamma_C)^N)$ we have that the second integral is bounded below by

$$-c - c\int_0^T \left(\int_{\Gamma_C} (u_n - w - g)_+^p \, d\Gamma\right)^{1/p} dt$$

$$\geq -c_\eta - \eta \int_0^T \int_0^t |v_n(s) - w_t(s)|^p_{L^p(\Gamma_C)} \, ds dt \tag{4.19}$$

for $\eta > 0$. Next, we examine the term $\left\langle e^{-\lambda(\cdot)} P(\mathbf{u}, w), \mathbf{y} \right\rangle$. Let $h(r, \mathbf{x}) = \int_{g(\mathbf{x})}^r p(s - g(\mathbf{x}))\, ds$ and define $H : L^2(\Gamma_C) \to [0, \infty)$ by

$$H(u) = \int_{\Gamma_C} h(u, \mathbf{x}) \, d\Gamma. \tag{4.20}$$

Then

$$\frac{d}{dt} H(u_n - w) = \langle DH(u_n - w), v_n - w' \rangle$$

$$= \int_{\Gamma_C} p(u_n - w - g)\,(v_n - w')\, d\Gamma = \langle P(\mathbf{u}, w), \mathbf{v} \rangle - \int_{\Gamma_C} p(u_n - w - g) w' \, d\Gamma. \tag{4.21}$$

Therefore,

$$\left\langle e^{-\lambda(\cdot)} P(\mathbf{u}, w), \mathbf{y} \right\rangle = \int_0^T e^{-2\lambda t} \langle P(\mathbf{u}, w), \mathbf{v} \rangle \, dt$$

$$= \int_0^T e^{-2\lambda t} \frac{d}{dt} H(u_n - w)\, dt + \int_0^T \int_{\Gamma_C} p(u_n - w - g) w' \, d\Gamma \tag{4.22}$$

$$\geq H(u_n(T) - w(T)) e^{-2\lambda T} - H(u_{0\varepsilon n}) + 2\lambda \int_0^T H(\mathbf{u}) e^{-2\lambda t} \, dt,$$

due to the assumptions that $w' \geq 0$ and $p(\cdot) \geq 0$. Similarly,

$$\left\langle e^{-\lambda(\cdot)} A\mathbf{u}, \mathbf{y} \right\rangle = \frac{1}{2} \langle A\mathbf{u}(T), \mathbf{u}(T) \rangle \, e^{-2\lambda T}$$

$$- \frac{1}{2} \langle A\mathbf{u}_{0\varepsilon}, \mathbf{u}_{0\varepsilon} \rangle + \lambda \int_0^T \langle A\mathbf{u}, \mathbf{u} \rangle \, e^{-2\lambda t} \, dt. \tag{4.23}$$

It follows from (4.19), (4.22), and (4.23) that

$$\langle \mathcal{A}_\lambda \mathbf{y}, \mathbf{y} \rangle \geq \delta^2 ||\mathbf{y}||^2_{L^2(0,T;W)} + \varepsilon e^{-2\lambda T} ||\gamma \mathbf{y}||^q_{L^q(0,T;(L^q(\Gamma_C))^N)}$$

$$- c_\eta - \eta \int_0^T \int_0^t |v_n(s) - w'(s)|^p_{L^p(\Gamma_C)} \, ds dt - H(u_{0\varepsilon n}).$$

We conclude that $\mathcal{A}_\lambda$ is coercive when η is sufficiently small and by Lemma 4.2 that $\mathcal{A}_\lambda : \mathcal{V}_q \to \mathcal{V}_q'$ is bounded. All the assumptions of Theorem 2.6 are satisfied now, and the proof of Theorem 4.1 is complete.

We use this result in the following section. However, we note that the theorem has merit of its own.

5. Existence and uniqueness. We obtain a solution for problem $\mathcal{P}$, when w is a known function, by deriving estimates on the solutions of $\mathcal{P}(\varepsilon)$ and passing to the limit $\varepsilon \to 0$, thus proving Theorem 3.1. We are still assuming that μ is Lipschitz continuous.

The proof of Theorem 3.1 is accomplished in a number of steps. We denote by c a generic positive constant which is independent of ε. Multiplying both sides of (4.2) by $\mathbf{v}\chi_{[0,t]}$ and using the above formulas along with the assumption that $w' \geq 0$, and performing routine manipulations, we obtain the following estimates for $\mathbf{v}^* \in Q\mathbf{v}$:

$$\begin{aligned}
&\frac{1}{2}|\mathbf{v}(t)|_H^2 - \frac{1}{2}|\mathbf{v}_0|_H^2 + \delta^2 \int_0^t ||\mathbf{v}||_W^2 \, ds + \frac{1}{2}\langle A\mathbf{u}(t), \mathbf{u}(t)\rangle \\
&\quad + \varepsilon \int_0^t \int_{\Gamma_C} |\gamma\mathbf{v}|^q \, d\Gamma ds + \langle \mathbf{v}^*, \mathbf{v}\chi_{[0,t]}\rangle + H(u_n(t) - w(t)) - H(u_{0\varepsilon n}) \\
&\qquad \leq \int_0^t \langle \mathbf{f}(s), \mathbf{v}(s)\rangle \, ds + \frac{1}{2}\langle A\mathbf{u}_{0\varepsilon}, \mathbf{u}_{0\varepsilon}\rangle. \qquad (5.1)
\end{aligned}$$

Now, when $\lambda = 0$ in (4.18), we obtain

$$\langle \mathbf{v}^*, \mathbf{v}\chi_{[0,t]}\rangle \geq -c - c\int_0^t \int_{\Gamma_C} (u_n - w - g)_+^p \, d\Gamma,$$

thus

$$\begin{aligned}
&\frac{1}{2}|\mathbf{v}(t)|_H^2 + \delta^2 \int_0^t ||\mathbf{v}||_W^2 \, ds + \frac{1}{2}\langle A\mathbf{u}(t), \mathbf{u}(t)\rangle + \varepsilon \int_0^t \int_{\Gamma_C} |\gamma\mathbf{v}|^q \, d\Gamma ds \\
&+ \int_{\Gamma_C} h(u_n(t,\mathbf{x}) - w(t,\mathbf{x}), \mathbf{x}) \, d\Gamma \leq c + \frac{1}{2}|\mathbf{v}_0|_H^2 + \frac{1}{2}\langle A\mathbf{u}_{0\varepsilon}, \mathbf{u}_{0\varepsilon}\rangle + H(u_{0\varepsilon n}) \\
&+ \frac{1}{2\delta^2}\int_0^t ||\mathbf{f}(s)||_{W'}^2 \, ds + \frac{\delta^2}{2}\int_0^t ||\mathbf{v}(s)||_W^2 \, ds + c\int_0^t \int_{\Gamma_C} (u_n - w - g)_+^p \, d\Gamma ds. \qquad (5.2)
\end{aligned}$$

The assumptions on $p(\cdot)$ given in (3.13) imply that if $r \geq g(\mathbf{x})$, then

$$h(r, \mathbf{x}) \geq \int_{g(\mathbf{x})}^r (\delta^2 (s - g)_+^{p-1} - c) \, ds = \frac{\delta^2}{p}(r - g(\mathbf{x}))_+^p - c(r - g(\mathbf{x}))_+. \qquad (5.3)$$

Now, since $p(r) = 0$ for $r \leq 0$, (5.3) holds also when $r < g(\mathbf{x})$. Then (5.2) yields

$$\begin{aligned}
&|\mathbf{v}(t)|_H^2 + \delta^2 \int_0^t ||\mathbf{v}||_W^2 \, ds + \langle A\mathbf{u}(t), \mathbf{u}(t)\rangle + 2\varepsilon \int_0^t \int_{\Gamma_C} |\gamma\mathbf{v}|^q \, d\Gamma ds \\
&\qquad + \frac{2\delta^2}{p}\int_{\Gamma_C} (u_n(t) - w(t) - g)_+^p d\Gamma - 2c\int_{\Gamma_C} (u_n(t) - w(t) - g)_+ \, d\Gamma \\
&\quad \leq c + |\mathbf{v}_0|^2 + \langle A\mathbf{u}_{0\varepsilon}, \mathbf{u}_{0\varepsilon}\rangle + 2H(\mathbf{u}_{0\varepsilon n}) + \frac{1}{\delta^2}\int_0^t ||\mathbf{f}(s)||_{W'}^2 \, ds \\
&\qquad + c\int_0^t \int_{\Gamma_C} (u_n - w - g)_+^p \, d\Gamma ds. \qquad (5.4)
\end{aligned}$$

Applying the Hölder inequality to the sixth term on the right-hand side we obtain

$$
\begin{aligned}
|\mathbf{v}(t)|_H^2 + \delta^2 \int_0^t ||\mathbf{v}||_W^2 \, ds + \langle A\mathbf{u}(t), \mathbf{u}(t)\rangle + 2\varepsilon \int_0^t \int_{\Gamma_C} |\gamma \mathbf{v}|^q \, d\Gamma ds \\
+ \ \frac{\delta^2}{p} \int_{\Gamma_C} (u_n(t) - w - g)_+^p \, d\Gamma \le c + |\mathbf{v}_0|^2 + \langle A\mathbf{u}_{0\varepsilon}, \mathbf{u}_{0\varepsilon}\rangle + 2H(\mathbf{u}_{0\varepsilon}) \\
+ \ \frac{1}{\delta^2} \int_0^t ||\mathbf{f}(s)||_{W'}^2 \, ds + c \int_0^t \int_{\Gamma_C} (u_n - w - g)_+^p \, d\Gamma ds. \qquad (5.5)
\end{aligned}
$$

Now using the Gronwall inequality yields

$$
\begin{aligned}
|\mathbf{v}(t)|_H^2 + \int_0^t ||\mathbf{v}||_W^2 ds + \langle A\mathbf{u}(t), \mathbf{u}(t)\rangle + \varepsilon \int_0^t \int_{\Gamma_C} |\gamma \mathbf{v}|^q \, d\Gamma ds \\
+ \int_{\Gamma_C} (u_n(t) - w - g)_+^p \, d\Gamma \le c, \qquad (5.6)
\end{aligned}
$$

where c does not depend on ε, q (for $q > p$) or w. If $\mathbf{w} \in \mathcal{V}_q$, then (5.6) and the definition of J imply

$$
\begin{aligned}
|\langle \varepsilon J\mathbf{v}, \mathbf{w}\rangle| &\le \varepsilon \langle J\mathbf{v}, \mathbf{v}\rangle^{(1/q')} \langle J\mathbf{w}, \mathbf{w}\rangle^{(1/q)} \\
&\le (\varepsilon \langle J\mathbf{v}, \mathbf{v}\rangle)^{(1/q')} \varepsilon^{(1/q)} ||\mathbf{w}||_{\mathcal{V}_q} \le c\varepsilon^{(1/q)} ||\mathbf{w}||_{\mathcal{V}_q}. \qquad (5.7)
\end{aligned}
$$

Thus, when $\mathbf{v}_\varepsilon$ is a solution of problem $\mathcal{P}(\varepsilon)$ we have

$$
\varepsilon J\mathbf{v}_\varepsilon \to 0 \quad \text{in } \mathcal{V}_q'. \qquad (5.8)
$$

From (5.6) and the growth conditions for $p(\cdot)$ we find that $Q(\mathbf{v}_\varepsilon, w)$ and $P(\mathbf{u}_\varepsilon, w)$ are bounded in $\mathcal{V}_p' \subseteq \mathcal{V}_q'$. Using Theorems 2.2 and 2.3 we find that there exists a subsequence, still denoted by $\varepsilon \to 0$, such that

$$
\begin{aligned}
\mathbf{v}_\varepsilon &\to \mathbf{v} \quad \text{weakly in } L^2(0,T;W), \qquad (5.9)\\
\mathbf{v}_\varepsilon' &\to \mathbf{v}' \quad \text{in } \mathcal{V}_q', \qquad (5.10)\\
\mathbf{u}_\varepsilon &\to \mathbf{u} \quad \text{in } C(0,T;U), \qquad (5.11)\\
\mathbf{v}_\varepsilon &\to \mathbf{v} \quad \text{in } L^2(0,T;U), \qquad (5.12)\\
M\mathbf{v}_\varepsilon &\to M\mathbf{v} \quad \text{weakly in } L^2(0,T;W'), \qquad (5.13)\\
A\mathbf{u}_\varepsilon &\to A\mathbf{u} \quad \text{weakly in } L^2(0,T;W'). \qquad (5.14)
\end{aligned}
$$

Here U denotes a space containing W with compact identity map and such that the trace map $\gamma : U \to L^2(\Gamma_C)^N$ is continuous. Letting $\mathbf{z}_\varepsilon$ be as in (4.5) and (4.6), (5.11) and (5.12) imply that, for a subsequence,

$$
\begin{aligned}
\widetilde{\gamma \mathbf{u}}_\varepsilon(\mathbf{x},t) &\to \widetilde{\gamma \mathbf{u}}(\mathbf{x},t) \quad \text{a.e. in } \Gamma_C \times (0,T), \qquad (5.15)\\
\widetilde{\gamma \mathbf{v}}_\varepsilon(\mathbf{x},t) &\to \widetilde{\gamma \mathbf{v}}(\mathbf{x},t) \text{ a.e. in } \Gamma_C \times (0,T), \qquad (5.16)\\
\mu\left(|\mathbf{v}_{\varepsilon T} - \ \mathbf{v}_*|\right) p_R\left(u_{\varepsilon n} - w - g\right) \mathbf{z}_\varepsilon &\rightharpoonup \xi \\
&\text{weakly in } L^{p'}(0,T;L^{p'}(\Gamma_C)^N). \qquad (5.17)
\end{aligned}
$$

LEMMA 5.1. *$P(\mathbf{u}_\varepsilon, w) \to P(\mathbf{u}, w)$ in $\mathcal{V}_q'$ and $P(\mathbf{u}_\varepsilon, w) \rightharpoonup P(\mathbf{u}, w)$ weakly in $\mathcal{V}_p'$.*

Proof. Let $\mathbf{w} \in \mathcal{V}_q$; then, by (3.13),

$$\begin{aligned}|\langle P(\mathbf{u}_\varepsilon, w) - P(\mathbf{u}, w), \mathbf{w}\rangle| &\leq \int_0^T \int_{\Gamma_C} K(1 + (u_{\varepsilon n} - w - g)_+^{p-2} + (u_n - w - g)_+^{p-2}) \\ &\quad \times |(u_{\varepsilon n} - w - g)_+ - (u_n - w - g)_+|\, |w_n|\, d\Gamma dt \\ &\leq c \int_0^T \left(\int_{\Gamma_C} (1 + (u_{\varepsilon n} - w - g)_+^p + (u_n - w - g)_+^p)\, d\Gamma \right)^{\frac{p-2}{p}} \\ &\quad \times \left(\int_{\Gamma_C} |(u_{\varepsilon n} - w - g)_+ - (u_n - w - g)_+|^r\, d\Gamma \right)^{\frac{1}{r}} \cdot \left(\int_{\Gamma_C} |w_n|^q d\Gamma \right)^{\frac{1}{q}} dt,\end{aligned}$$

where $r = pq(2q-p)^{-1}$. It follows from (5.6) that

$$\begin{aligned}&|\langle P(\mathbf{u}_\varepsilon, w) - P(\mathbf{u}, w), \mathbf{w}\rangle| \\ &\leq c \left(\int_0^T \int_{\Gamma_C} |(u_{\varepsilon n} - w - g)_+ - (u_n - w - g)_+|^r\, d\Gamma dt \right)^{\frac{1}{r}} ||\mathbf{w}||_{\mathcal{V}_q}. \end{aligned} \tag{5.18}$$

Now note that $r < p$ and so estimate (5.6) implies the functions $|(u_\varepsilon - w - g)_+ - (u_n - w - g)_+|^r$ are uniformly integrable. Then (5.15) and the Vitali convergence theorem imply

$$\lim_{\varepsilon \to 0} \int_0^T \int_{\Gamma_C} |(u_{\varepsilon n} - w - g)_+ - (u_n - w - g)_+|^r d\Gamma dt = 0.$$

Now

$$\begin{aligned}&\|P(\mathbf{u}_\varepsilon, w) - P(\mathbf{u}, w)\|_{\mathcal{V}_q'} \\ &\quad \leq c \left(\int_0^T \int_{\Gamma_C} |(u_{\varepsilon n} - w - g)_+ - (u_n - w - g)_+|^r\, d\Gamma\, dt \right)^{\frac{1}{r}},\end{aligned}$$

and hence $P(\mathbf{u}_\varepsilon, w) \to P(\mathbf{u}, w)$ in $\mathcal{V}_q'$.

To obtain the other assertion, we note that $P(\mathbf{u}_\varepsilon, w)$ is bounded in $\mathcal{V}_p'$, and therefore it has a convergent subsequence such that $P(\mathbf{u}_\varepsilon, w) \rightharpoonup \ell$ weakly in $\mathcal{V}_p'$. However, $\mathcal{V}_q$ is dense in $\mathcal{V}_p$ and so $\ell = P(\mathbf{u}, w)$. Since this holds for every weakly convergent subsequence, it follows that $P(\mathbf{u}_\varepsilon, w) \rightharpoonup P(\mathbf{u}, w)$.

LEMMA 5.2. *For each* $\mathbf{w} \in \mathcal{V}_p$,

$$\langle \gamma_T^* \xi, \mathbf{w}\rangle \leq \int_0^T \int_{\Gamma_C} \mu p_R \left(|\mathbf{v}_T - \mathbf{v}_* + \ \mathbf{w}_T| - |\mathbf{v}_T - \mathbf{v}_*| \right) d\Gamma\, dt, \tag{5.19}$$

where $\mu(|\mathbf{v}_T - \mathbf{v}_*|)$ *and* $p_R = p_R(u_n - w - g)$.

Proof. To simplify the notation we let $F = p_R(u_n - w - g)$, $F_\varepsilon = p_R(u_{\varepsilon n} - w - g)$, $\mu = \mu(|\mathbf{v}_T - \mathbf{v}_*|)$, and $\mu_\varepsilon = \mu(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_*|)$. First suppose that $\mathbf{w} \in \mathcal{V}_q$. It follows from the assumptions on $\mathbf{z}_\varepsilon$ that $\mathbf{z}_\varepsilon \cdot \mathbf{w}_T \leq (|\mathbf{v}_{\varepsilon T} - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_{\varepsilon T} - \mathbf{v}_*|)$ for a.e. t and a.e. $\mathbf{x}$. Therefore,

$$\begin{aligned}\langle \gamma_T^* \xi, \mathbf{w}\rangle &= \lim_{\varepsilon \to 0} \int_0^T \int_{\Gamma_C} F_\varepsilon \mu_\varepsilon \mathbf{z}_\varepsilon \cdot \mathbf{w}_T\, d\Gamma dt \\ &\leq \liminf_{\varepsilon \to 0} \int_0^T \int_{\Gamma_C} F_\varepsilon \mu_\varepsilon \left(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_{\varepsilon T} - \mathbf{v}_*| \right) d\Gamma dt. \end{aligned} \tag{5.20}$$

Now the integrand converges pointwise to $F\mu\left(|\mathbf{v}_T-\mathbf{v}_*+\mathbf{w}_T|-|\mathbf{v}_T-\ \mathbf{v}_*|\right)$ and is bounded in absolute value by $c(1+(u_{\varepsilon n}-w-g)_+^{p-1})|\mathbf{w}_T|$. These functions are bounded in $L^r((0,T)\times\Gamma_C)$, independently of ε, where $r\equiv pq/(pq+p-q)$. Indeed, $(p-1)rq/(q-r)=p$, and thus

$$\begin{aligned}(u_{\varepsilon n}-w-g)_+^{(p-1)r}\,|\mathbf{w}_T|^r &\le (u_{\varepsilon n}-w-g)_+^{\frac{(p-1)rq}{q-r}}+|\mathbf{w}_T|^q\\ &=(u_{\varepsilon n}-w-g)_+^{p}+|\mathbf{w}_T|^q\,,\end{aligned}$$

which is bounded in L^1, independent of ε. Therefore, using the Vitali convergence theorem in (5.20), we may pass to the limit and obtain (5.19) for all $\mathbf{w}\in\mathcal{V}_q$, and since $\mathcal{V}_q$ is dense in $\mathcal{V}_p$ this inequality holds for all $\mathbf{w}\in\mathcal{V}_p$. This proves the lemma.

Next, from (4.2), (5.13), (5.14), (5.9), and Lemma 5.1 we obtain

$$\mathbf{v}'+M\mathbf{v}+A\mathbf{u}+\gamma_T^*\xi+P\left(\mathbf{u},w\right)=\mathbf{f}\ \text{in}\ \mathcal{V}_q'.$$

Since $\gamma_T^*\xi,\ A\mathbf{u},\ M\mathbf{v}$ and $\mathbf{f}$ are all in $\mathcal{V}_p'$, so is $\mathbf{v}'$. This proves the existence part of the theorem.

Proof of uniqueness. Suppose $\mathbf{v}_1$ and $\mathbf{v}_2$ are two solutions of $\mathcal{P}$. Let, for $i=1,2,\ \mathbf{u}_i(t)=\mathbf{u}_0+\int_0^t\mathbf{v}_i(s)\,ds$. It follows that

$$\begin{aligned}&\frac{1}{2}|\mathbf{v}_1(t)-\mathbf{v}_2(t)|_H^2+\int_0^t\langle M\mathbf{v}_1-M\mathbf{v}_2,\mathbf{v}_1-\mathbf{v}_2\rangle\,ds\\ &\quad+\int_0^t\langle A(\mathbf{u}_1-\mathbf{u}_2),\mathbf{v}_1-\mathbf{v}_2\rangle\,ds+\int_0^t\langle\gamma_T^*\xi_1-\gamma_T^*\xi_2,\mathbf{v}_1-\mathbf{v}_2\rangle\,ds\\ &\quad+\int_0^t\langle P\left(\mathbf{u}_1,w\right)-P\left(\mathbf{u}_2,w\right),\mathbf{v}_1-\mathbf{v}_2\rangle\,ds=0.\end{aligned}\tag{5.21}$$

Thus, if we denote by c a positive generic constant, we have

$$\begin{aligned}&\frac{1}{2}|\mathbf{v}_1(t)-\mathbf{v}_2(t)|_H^2+\frac{1}{2}\left\langle A(\mathbf{u}_1(t)-\mathbf{u}_2(t)),\mathbf{u}_1(t)-\mathbf{u}_2(t)\right\rangle\\ &+\int_0^t\left\langle P\left(\mathbf{u}_1,w\right)-P\left(\mathbf{u}_2,w\right),\mathbf{v}_1-\mathbf{v}_2\right\rangle\,ds+\delta^2\int_0^t||\mathbf{v}_1-\mathbf{v}_2||_W^2\,ds\\ &+\int_0^t\langle\gamma_T^*\xi_1-\gamma_T^*\xi_2,\mathbf{v}_1-\mathbf{v}_2\rangle\,ds\le c\int_0^t|\mathbf{v}_1(s)-\mathbf{v}_2(s)|_H^2\,ds.\end{aligned}\tag{5.22}$$

Let $F^i=p_{_R}(u_{in}-w-g)$, $\mu^i=\mu(|\mathbf{v}_{iT}-\mathbf{v}_*|)$, for $i=1,2$; then using condition (5.20) we observe

$$\begin{aligned}&\int_0^t\langle\gamma_T^*\xi_1-\gamma_T^*\xi_2,\mathbf{v}_1-\mathbf{v}_2\rangle\,ds\\ &\quad\ge\int_0^t\int_{\Gamma_C}\left(F^1\mu^1-F^2\mu^2\right)\left(|\mathbf{v}_{1T}-\mathbf{v}_*|-|\mathbf{v}_{2T}-\mathbf{v}_*|\right)\,d\Gamma ds.\end{aligned}$$

Consequently, the last term on the left-hand side in (5.22) dominates

$$-c\int_0^t\int_{\Gamma_C}F^2\,|\mathbf{v}_{1T}-\mathbf{v}_{2T}|^2\,d\Gamma ds-c\int_0^t\int_{\Gamma_C}|F^1-F^2|\,|\mathbf{v}_{1T}-\mathbf{v}_{2T}|\,d\Gamma ds.\tag{5.23}$$

The third term in (5.22) is greater than or equal to

$$-\int_0^t \int_{\Gamma_C} |p(u_{1n}-w-g)-p(u_{2n}-w-g)|\,|v_{1n}-v_{2n}|\,d\Gamma ds. \tag{5.24}$$

From the assumptions on $p(\cdot)$ and from (5.22) we obtain

$$\begin{aligned}
&|\mathbf{v}_1(t)-\mathbf{v}_2(t)|_H^2 + \langle A(\mathbf{u}_1(t)-\mathbf{u}_2(t)), \mathbf{u}_1(t)-\mathbf{u}_2(t)\rangle + \delta^2 \int_0^t ||\mathbf{v}_1-\mathbf{v}_2||_W^2\, ds \\
&\quad \le c\int_0^t \int_{\Gamma_C} (1+|(u_{1n}-w-g)_+|^2+|(u_{2n}-w-g)_+|^2) \\
&\qquad\qquad \times |u_{1n}-u_{2n}|\,|v_{1n}-v_{2n}|\,d\Gamma ds \\
&\quad + c\int_0^t |\mathbf{v}_1(s)-\mathbf{v}_2(s)|_H^2\, ds + c\int_0^t \int_{\Gamma_C} |\mathbf{v}_{1T}-\mathbf{v}_{2T}|^2\, d\Gamma ds.
\end{aligned}$$

Since $(u_n-w-g)_+ \in L^\infty(0,T;L^p(\Gamma_C))$, we obtain, with another c which depends on $\mathbf{u}_1$ and $\mathbf{u}_2$,

$$\begin{aligned}
|\mathbf{v}_1(t)-\mathbf{v}_2(t)|_H^2 + \delta^2 \int_0^t ||\mathbf{v}_1-\mathbf{v}_2||_W^2\, ds &\le c\int_0^t ||\mathbf{v}_1-\mathbf{v}_2||_U^2\, dt \\
&+ c\int_0^t \left(\int_{\Gamma_C} |u_{1n}-u_{2n}|^4 d\Gamma\right)^{\frac14} \left(\int_{\Gamma_C} |v_{1n}-v_{2n}|^4 d\Gamma\right)^{\frac14} ds \\
&+ c\int_0^t |\mathbf{v}_1(s)-\mathbf{v}_2(s)|_H^2\, ds \\
&\le c\int_0^t ||\mathbf{u}_1-\mathbf{u}_2||_W\, ||\mathbf{v}_1-\mathbf{v}_2||_W\, ds + c\int_0^t |\mathbf{v}_1(s)-\mathbf{v}_2(s)|_H^2\, ds \\
&+ K\int_0^t ||\mathbf{v}_1-\mathbf{v}_2||_U^2\, dt,
\end{aligned}$$

where we used the fact that the trace map $W \to L^4(\partial\Omega)$ is continuous. It follows from the compactness of the embedding $U \to W$ that

$$\begin{aligned}
|\mathbf{v}_1(t)-\mathbf{v}_2(t)|_H^2 + \frac{\delta^2}{2}\int_0^t ||\mathbf{v}_1-\mathbf{v}_2||_W^2\, ds &\le c_{\delta T}\int_0^t \int_0^s ||\mathbf{v}_1-\mathbf{v}_2||_W^2\, dr ds + K_\varepsilon \int_0^t |\mathbf{v}_1(s)-\mathbf{v}_2(s)|_H^2\, ds \\
&+ \varepsilon \int_0^t ||\mathbf{v}_1-\mathbf{v}_2||_W^2\, dt.
\end{aligned}$$

Choosing $\varepsilon = \frac{\delta^2}{4}$ and adjusting the constants yields

$$\begin{aligned}
|\mathbf{v}_1(t)-\mathbf{v}_2(t)|_H^2 + \frac{\delta^2}{4}\int_0^t ||\mathbf{v}_1-\mathbf{v}_2||_W^2\, ds &\le c_{\delta T}\int_0^t \left(\int_0^s ||\mathbf{v}_1-\mathbf{v}_2||_W^2\, dr + |\mathbf{v}_1(s)-\mathbf{v}_2(s)|_H^2\right) ds.
\end{aligned}$$

By the Gronwall inequality we obtain $\mathbf{v}_1 = \mathbf{v}_2$. This concludes the proof of Theorem 3.1 in the case that p satisfies (3.13).

In the case when $p(\cdot)$ satisfies (3.12) the proof is much easier, not requiring the consideration of the approximate problems where εJ was added in.

THEOREM 5.3. *Let $p \geq 2$ and let $w \in L^p(0,T;L^p(\Gamma_C))$, $w' \in L^p(0,T;L^p(\Gamma_C))$, $w' \geq 0$, $\mathbf{u}_0 \in V_p$, $\mathbf{v}_0 \in H$, $\mathbf{f} \in \mathcal{V}_p'$ and assume $\mu^*(r) = \mu_c(r)$, where μ_c is bounded and Lipschitz. Then there exists $\xi \in L^{p'}(0,T;L^{p'}(\Gamma_C)^N)$ and $\mathbf{v} \in L^2(0,T;W)$ such that*

$$(u_n - w - g)_+ \in L^\infty(0,T;L^p(\Gamma_C)), \tag{5.25}$$

$$\mathbf{v}' + M\mathbf{v} + A\mathbf{u} + P(\mathbf{u},w) + \gamma_T^*\xi = \mathbf{f} \ \text{ in } \ \mathcal{V}_p', \tag{5.26}$$

$$\mathbf{v}(0) = \mathbf{v}_0\,, \ \mathbf{u}(t) = \mathbf{u}_0 + \int_0^t \mathbf{v}(s)\,ds, \tag{5.27}$$

and

$$\langle \gamma_T^*\xi, \mathbf{w}\rangle \leq \int_0^T \int_{\Gamma_C} \mu p_{_R} \left(|\mathbf{v}_T - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_T - \mathbf{v}_*|\right) d\Gamma\, dt, \tag{5.28}$$

where $\mu = \mu(|\mathbf{v}_T - \mathbf{v}_*|)$ and $p_{_R} = p_{_R}(u_n - w - g)$.

Moreover, if the function $p(\cdot)$ satisfies (3.12)*, the solution $\{\mathbf{u},\mathbf{v}\}$ is unique.*

We note that (3.28) and the fact that $\mathbf{v}_1 = \mathbf{v}_2$ imply $\gamma_T^*\xi_1 = \gamma_T^*\xi_2$; however, we do not know if $\xi_1 = \xi_2$.

6. Discontinuous friction coefficient. In this section we consider the case when the coefficient of friction is a discontinuous function of the slip speed and establish Theorem 3.2. This is the case often described in elementary courses where it is stated that the coefficient of sliding friction is smaller than the coefficient of static friction. Therefore, we assume that the function μ has a jump discontinuity at zero, becoming smaller when slip takes place, and is represented by the friction graph μ^* (3.5).

To investigate this case when p satisfies (3.13), we regularize the graph μ^* by defining $\mu_c(r) = \mu_d$ for all $r \leq 0$ and

$$\mu_\varepsilon(r) = \mu_c(r) - h_\varepsilon'(r) + \eta,$$

where $2\eta = \mu_s - \mu_d$ and $h_\varepsilon(r) \equiv (\eta^2 r^2 + \varepsilon)^{1/2}$, for $0 < \varepsilon$ small. Thus, η is half the size of the jump at 0 between μ_d and μ_s. From this definition, it follows that

$$\lim_{\varepsilon\to 0} \mu_\varepsilon(r) = \begin{cases} \mu_c(r) \quad \text{if } \ r > 0, \\ \mu_c(r) + 2\eta = \mu_s \quad \text{if } \ r < 0, \\ \mu_d + \eta \text{ if } r = 0 \end{cases}$$

which is a function whose graph has a jump of height $2\eta = \mu_s - \mu_d$ at $r = 0$.

Let $\mathbf{v}_\varepsilon$ be the solution of the approximate problem (4.2)–(4.6) in which μ is replaced with μ_ε. Then, estimate (5.6) holds for $\mathbf{v}_\varepsilon$ and, consequently, there exists a subsequence such that (5.9)–(5.17) hold. Passing to a further subsequence if necessary, we may assume there exists $\psi \in L^\infty(0,T;L^\infty(\Gamma_C))$ such that

$$h_\varepsilon'(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_*|) \to \psi \text{ weak} * \text{ in } L^\infty(0,T;L^\infty(\Gamma_C))\,.$$

We note that Lemma 5.1 still holds. As above, we let $F = p_{_R}(u_n - w - g)$ and

$F_\varepsilon = p_{_R}(u_{\varepsilon n} - w - g)$. Let $\mathbf{w} \in \mathcal{V}_q$,

$$\begin{aligned}
\langle \gamma_T^* \xi, \mathbf{w} \rangle &= \lim_{\varepsilon \to 0} \int_0^T \int_{\Gamma_C} F_\varepsilon \mu_\varepsilon \mathbf{z}_\varepsilon \cdot \mathbf{w}_T \, d\Gamma dt \\
&\leq \liminf_{\varepsilon \to 0} \int_0^T \int_{\Gamma_C} F_\varepsilon \mu_\varepsilon \left(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_{\varepsilon T} - \mathbf{v}_*| \right) d\Gamma dt \\
&= \liminf_{\varepsilon \to 0} \Bigg[\int_0^T \int_{\Gamma_C} F_\varepsilon \left(\mu_c \left(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_*| \right) - \psi + \eta \right) \\
&\qquad \times \left(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_{\varepsilon T} - \mathbf{v}_*| \right) d\Gamma dt \\
&\qquad + \int_0^T \int_{\Gamma_C} F_\varepsilon \left(\psi - h_\varepsilon' \left(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_*| \right) \right) \\
&\qquad \times \left(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_{\varepsilon T} - \mathbf{v}_*| \right) d\Gamma dt \Bigg].
\end{aligned}$$

As in the proof of Lemma 5.2, the first integral on the right-hand side converges to

$$\int_0^T \int_{\Gamma_C} \left(\mu_c \left(|\mathbf{v}_T - \mathbf{v}_*| \right) - \psi + \eta \right) p_{_R} \left(|\mathbf{v}_T - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_T - \mathbf{v}_*| \right) d\Gamma dt,$$

where $p_{_R} = p_{_R}\left(u_n - w - g\right)$. We need to show that the second integral converges to zero. This follows from the observation that, since $p_{_R}$ is bounded,

$$|p_{_R}(u_{\varepsilon n} - w - g)(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_* + \mathbf{w}_T| - |\ \mathbf{v}_{\varepsilon T} - \mathbf{v}_*|)|$$

is bounded in $L^2((0,T) \times \Gamma_C)$, independently of ε, and converges pointwise to $|F\left(|\mathbf{v}_T - \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_T - \mathbf{v}_*|\right)|$, which lies in $L^2((0,T) \times \Gamma_C)$. Thus, the sequence is uniformly integrable, and by the Vitali convergence theorem it converges strongly in $L^1((0,T) \times \Gamma_C)$. Since $\psi - h_\varepsilon'(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_*|)$ converges weak$*$ in L^∞ to zero, the second integral converges to zero as desired. Next, we consider ψ.

First, note that, from the convexity of h_ε,

$$h_\varepsilon' \left(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_*| \right) z \leq h_\varepsilon \left(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_*| + z \right) - h_\varepsilon \left(|\mathbf{v}_{\varepsilon T} - \mathbf{v}_*| \right),$$

thus for arbitrary $z \in L^1\left(0, T; L^1\left(\Gamma_C\right)\right)$

$$\int_0^T \int_{\Gamma_C} \psi z d\Gamma dt \leq \int_0^T \int_{\Gamma_C} |\eta \left(|\mathbf{v}_T - \mathbf{v}_*| + z \right)| - |\eta(|\mathbf{v}_T - \mathbf{v}_*|)| \, d\Gamma dt$$

which implies that, for a.e. t,

$$\psi z \leq |\eta \left(|\mathbf{v}_T - \mathbf{v}_*| + z \right)| - |\eta(|\mathbf{v}_T - \mathbf{v}_*|)|$$

for a.e. $\mathbf{x}$. Letting $\theta\left(r\right) \equiv |\eta r|$, it follows that, for a.e. $\mathbf{x}, t$,

$$\psi\left(t, \mathbf{x}\right) \in \partial\theta \left(|\mathbf{v}_T - \mathbf{v}_*| \left(t, \mathbf{x}\right) \right).$$

Therefore, for a.e. $t, \mathbf{x}$,

$$\psi\left(t, \mathbf{x}\right) \in \left[-\eta, \eta\right].$$

More particularly, if $|\mathbf{v}_T - \mathbf{v}_*| > 0$, $\psi = \eta$, while if $|\mathbf{v}_T - \mathbf{v}_*| = 0$, the above holds. Therefore, the pair

$$\left(|\mathbf{v}_T - \mathbf{v}_*|,\ \mu_c\left(|\mathbf{v}_T - \mathbf{v}_*|\right) - \psi + \eta\right)$$

is an element of the graph of μ^*, a.e. The proof of Theorem 3.2 is now complete in the case where $p(\cdot)$ satisfies (3.13). Theorem 3.2 holds in the case where $p(\cdot)$ satisfies (3.12) from arguments similar to the above but without the necessity of dealing with the limit as $\varepsilon \to 0$ in the solutions of the approximate problems in which $\varepsilon J\mathbf{v}_\varepsilon$ was added.

The uniqueness of the solution remains an open question.

7. Dependence on w. In this section we investigate the dependence of the solution of (3.27)–(3.30) on w in the situation of (3.12) and $\mu^* = \mu = \mu_c$. Therefore, in this section we do not need to employ the truncation $p_{_R}$. We need to identify the dependence of $\gamma_T^*\xi$ on w and for this reason we write $\gamma_T^*\xi_w$ and rewrite (3.27)–(3.29) as follows:

$$\mathbf{v}' + M\mathbf{v} + A\mathbf{u} + \gamma_T^*\xi_w + P(\mathbf{u}, w) = \mathbf{f} \text{ in } \mathcal{V}_2', \tag{7.1}$$

$$\mathbf{v}(0) = \mathbf{v}_0,\ \mathbf{u}(t) = \mathbf{u}_0 + \int_0^t \mathbf{v}(s)\ ds, \tag{7.2}$$

and

$$\langle \gamma_T^*\xi_w, \mathbf{w}\rangle \le \int_0^T \int_{\Gamma_C} \mu p\left(|\mathbf{v}_T - \ \mathbf{v}_* + \mathbf{w}_T| - |\mathbf{v}_T - \mathbf{v}_*|\right)\ d\Gamma dt, \tag{7.3}$$

where $\mu = \mu(|\mathbf{v}_T - \mathbf{v}_*|)$ and $p = p(u_n - w - g)$.

Now let w_i, for $i = 1, 2$, be two wear functions as above and let $\mathbf{v}^i$ denote the corresponding solutions of problem (7.1)–(7.3). We need the following estimates. From (7.3) we obtain

$$\begin{aligned}
&\int_0^t \langle \gamma_T^*\xi_{w_1} - \gamma_T^*\xi_{w_2}, \mathbf{v}^1 - \mathbf{v}^2\rangle\, ds \\
&\quad \ge -\int_0^t \int_{\Gamma_C} F^1\mu^1 \left(\left|\mathbf{v}_T^2 - \mathbf{v}_*\right| - \left|\mathbf{v}_T^1 - \mathbf{v}_*\right|\right)\ d\Gamma\, ds \\
&\qquad -\int_0^t \int_{\Gamma_C} F^2\mu^2 \left(\left|\mathbf{v}_T^1 - \mathbf{v}_*\right| - \left|\mathbf{v}_T^2 - \mathbf{v}_*\right|\right)\ d\Gamma\, ds \\
&\quad = \int_0^t \int_{\Gamma_C} \left(F^2\mu^2 - F^1\mu^1\right)\left(|\mathbf{v}_T^2 - \mathbf{v}_*| - |\mathbf{v}_T^1 - \mathbf{v}_*|\right)\ d\Gamma\, ds,
\end{aligned}$$

where $F^i = p(u_n^i - w_i - g)$ and $\mu^i = \mu(|\mathbf{v}_T^i - \mathbf{v}_*|)$, for $i = 1, 2$. Let c be a positive constant which depends on Lip_μ, Lip_p, $p(\cdot)$, and the bounds on μ and $p(\cdot)$; then

$$\int_0^t \langle \gamma_T^*\xi_{w_1} - \gamma_T^*\xi_{w_2}, \mathbf{v}^1 - \mathbf{v}^2\rangle ds \ge -c\int_0^t \int_{\Gamma_C} \left|\mathbf{v}_T^1 - \mathbf{v}_T^2\right|^2\ d\Gamma ds$$

$$-\ c\int_0^t \int_{\Gamma_C} \left|\mathbf{v}_T^2 - \mathbf{v}_T^1\right| |w_1 - w_2|\ d\Gamma d - c\int_0^t \int_{\Gamma_C} \left|\mathbf{v}_T^1 - \mathbf{v}_T^2\right| \left|u_n^1 - u_n^2\right|\ d\Gamma ds. \tag{7.4}$$

Next, we consider the term $\int_0^t \langle P\left(\mathbf{u}^1, w_1\right) - P\left(\mathbf{u}^2, w_2\right), \mathbf{v}^1 - \mathbf{v}^2\rangle \, ds$. From (3.11) and (3.20), the definition of $P(\mathbf{u}, w)$, we obtain that this expression is no smaller than

$$-\int_0^t \int_{\Gamma_C} \left(p\left(u_n^1 - w_1 - g\right) - p\left(u_n^2 - w_2 - g\right)\right)\left(v_n^1 - v_n^2\right) \, d\Gamma ds$$

$$\geq -c \int_0^t \int_{\Gamma_C} \left(\left|u_n^1 - u_n^2\right| + |w_1 - w_2|\right)\left|v_n^1 - v_n^2\right| \, d\Gamma ds. \tag{7.5}$$

Now let U be a space in which V_2 embeds compactly and for which the trace map from U to $L^2(\partial\Omega)$ is continuous. Then, after adjusting the constant c and denoting by H_C the Hilbert space $L^2(\Gamma_C)$, we obtain from (7.4) and (7.5)

$$\int_0^t \langle \gamma_T^* \xi_{w_1} - \gamma_T^* \xi_{w_2}, \mathbf{v}^1 - \mathbf{v}^2\rangle \, ds + \int_0^t \langle P\left(\mathbf{u}^1, w_1\right) - P\left(\mathbf{u}^2, w_2\right), \mathbf{v}^1 - \mathbf{v}^2\rangle \, ds$$

$$\geq -c \int_0^t \left|\left|\mathbf{v}^1 - \mathbf{v}^2\right|\right|_U^2 ds - c \int_0^t |w_1 - w_2|_{H_C}^2 \, ds. \tag{7.6}$$

It follows from (7.6) and (7.1) that

$$\begin{aligned}
&\left|\mathbf{v}^1(t) - \mathbf{v}^2(t)\right|_H^2 + \delta^2 \int_0^t \left|\left|\mathbf{v}^1(s) - \mathbf{v}^2(s)\right|\right|_{V_2}^2 ds \\
&\qquad + \frac{1}{2}\langle A\left(\mathbf{u}^1(t) - \mathbf{u}^2(t)\right), \mathbf{u}^1(t) - \mathbf{u}^2(t)\rangle \\
&\leq c \int_0^t \left|\left|\mathbf{v}^1 - \mathbf{v}^2\right|\right|_U^2 ds + c \int_0^t |w_1 - w_2|_{H_C}^2 \, ds \\
&\qquad + \delta^2 \int_0^t \left|\mathbf{v}^1(s) - \mathbf{v}^2(s)\right|_H^2 ds.
\end{aligned}$$

By the compactness of the embedding $V_2 \to U$ we have $||\mathbf{z}||_U^2 \leq \frac{\delta^2}{2} ||\mathbf{z}||_{V_2}^2 + c_\delta |\mathbf{z}|_H^2$; hence,

$$\left|\mathbf{v}^1(t) - \mathbf{v}^2(t)\right|_H^2 + \frac{\delta^2}{2} \int_0^t \left|\left|\mathbf{v}^1(s) - \mathbf{v}^2(s)\right|\right|_{V_2}^2 ds$$

$$\leq c_\delta \int_0^t \left|\mathbf{v}^1(s) - \mathbf{v}^2(s)\right|_H^2 ds + c \int_0^t |w_1 - w_2|_{H_C}^2 \, ds.$$

It follows from the Gronwall inequality that

$$\begin{aligned}
\left|\mathbf{v}^1(t) - \mathbf{v}^2(t)\right|_H^2 + &\int_0^t \left|\left|\mathbf{v}^1(s) - \quad \mathbf{v}^2(s)\right|\right|_{V_2}^2 ds \\
&\leq c(\delta, T) \int_0^t |w_1 - w_2|_{H_C}^2 \, ds,
\end{aligned} \tag{7.7}$$

where the constant c depends on the indicated quantities and the bounds and Lipschitz constants of p and μ but not on the choice of w_i. We conclude with the following theorem.

THEOREM 7.1. *The solutions* $\mathbf{v}$ *of problem* (3.27)–(3.30) *depend continuously on* w.

8. Archard law. We now consider Theorem 3.3. We use a fixed point argument to prove Theorem 3.3, which guarantees the existence and uniqueness of the weak solution. Since $p(\cdot)$ is assumed to be bounded, we do not need to employ the truncation p_R.

The Archard law of wear, in its differential form (3.9), may be written as

$$w' = \Psi(\mathbf{v}_T)\, p(u_n - w - g),$$

where $\Psi(\mathbf{v}_T) \equiv k_w \mu(|\mathbf{v}_T - \mathbf{v}_*|) s_c(|\mathbf{v}_T - \mathbf{v}_*|)$. It follows from our assumptions that Ψ is bounded, nonnegative, and Lipschitz continuous. Let $\mathbf{v}^i \in \mathcal{V}_2$ and w_i, for $i = 1, 2$, be the solutions of the problem

$$w_i, w_i' \in L^2(0, T; H_C), \tag{8.1}$$

$$w_i' = \Psi\left(\mathbf{v}_T^i\right) p\left(u_n^i - w_i - g\right), \tag{8.2}$$

$$w_i(\cdot, 0) = 0. \tag{8.3}$$

Since the function Ψ is bounded, we actually have

$$w, w' \in L^\infty(0, T; L^\infty(\Gamma_C)),$$

and so these functions may be considered as known wear functions in the preceding theory. Thus,

$$\begin{aligned}
&\frac{1}{2}\left|w_1(t) - w_2(t)\right|_{H_C}^2 \\
&\quad \le c(\Psi, R) \int_0^t \left(|u_n^1 - u_n^2|_{H_C} + |w_1 - w_2|_{H_C}\right)\left(|w_1 - w_2|_{H_C}\right) ds \\
&\qquad + c(\ \mathrm{Lip}_\Psi, R, p) \int_0^t \left|\mathbf{v}_T^1 - \mathbf{v}_T^2\right|_{H_C^N} |w_1 - w_2|_{H_C}\, ds,
\end{aligned}$$

where $H_C = L^2(\Gamma_C)$. It follows that

$$\begin{aligned}
&|w_1(t) - w_2(t)|_{H_C}^2 \\
&\quad \le c(\Psi, R, p,\ \mathrm{Lip}_\Psi, T)\left(\int_0^t |w_1 - w_2|_{H_C}^2\, ds + \int_0^t \left|\left|\mathbf{v}^2 - \mathbf{v}^1\right|\right|_U^2\, ds\right),
\end{aligned}$$

where U is an intermediate space. Thus, by the Gronwall inequality,

$$|w_1(t) - w_2(t)|_{H_C}^2 \le c(\Psi, R, p,\ \mathrm{Lip}_\Psi, T) \int_0^t \left|\left|\mathbf{v}^1 - \mathbf{v}^2\right|\right|_U^2\, ds. \tag{8.4}$$

Now we construct the following mapping. Starting with $\mathbf{v} \in \mathcal{V}_2$, we denote by $w(\mathbf{v})$ the solution of problem (8.1)–(8.3), with i omitted. Then we use $w(\mathbf{v})$ as the wear function in the system (7.1)–(7.3). In this manner we define a mapping, $\Lambda : \mathcal{V}_2 \to \mathcal{V}_2$, where $\mathbf{z} = \Lambda\mathbf{v}$, and $\mathbf{z}$ is the solution of (7.1)–(7.3) with the given wear function $w(\mathbf{v})$. Now, from (7.7) and (8.4) we obtain

$$\int_0^t \left|\left|\Lambda\mathbf{v}^1 - \Lambda\mathbf{v}^2\right|\right|_{V_2}^2\, ds \le c(\delta, T, \Psi, R, p,\ \mathrm{Lip}_\Psi) \int_0^t \int_0^s \left|\left|\mathbf{v}^1 - \mathbf{v}^2\right|\right|_{V_2}^2\, dr ds.$$

By iterating this inequality m times we find that every Λ^m is a contraction mapping on $\mathcal{V}_2$ for all sufficiently large m. Consequently, Λ has a unique fixed point, which is the unique solution of problem $\mathcal{P}$. This establishes Theorem 3.3.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] A. Amassad, M. Shillor, and M. Sofonea, *A quasistatic contact problem with slip dependent coefficient of friction*, Math. Methods Appl. Sci., 22 (1999), pp. 267–284.

[3] A. Amassad, K. L. Kuttler, M. Rochdi, and M. Shillor, *Existence for a Dynamic Thermoviscoelastic Contact Problem with Slip-Dependent Friction Coefficient*, preprint.

[4] L.-E. Andersson, *A quasistatic frictional problem with normal compliance*, Nonlinear Anal., 16 (1991), pp. 347–370.

[5] K. T. Andrews, K. L. Kuttler, and M. Shillor, *On the dynamic behaviour of a thermoviscoelastic body in frictional contact with a rigid obstacle*, European J. Appl. Math., 8 (1997), pp. 417–436.

[6] K. T. Andrews, A. Klarbring, M. Shillor, and S. Wright, *A dynamic contact problem with friction and wear*, Internat. J. Engrg. Sci., 35 (1997), pp. 1291–1309.

[7] M. Cocu, E. Pratt, and M. Raous, *Formulation and approximation of quasistatic frictional contact*, Internat. J. Engrg. Sci., 34 (1996), pp. 783–798.

[8] G. Duvaut and J. L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, New York, 1976.

[9] I. Figueirdo and L. Trabucho, *A class of contact and friction dynamic problems in thermoelasticity and in thermoviscoelasticity*, Internat. J. Engrg. Sci., 33 (1995), pp. 45–66.

[10] I. R. Ionescu and J.-C. Paumier, *On the contact problem with slip dependent friction in elastodynamics*, European J. Mech. A Solids, 13 (1994), pp. 555–568.

[11] J. Jarusek and Ch. Eck, *Dynamic contact problems with friction in linear viscoelasticity*, C. R. Acad. Sci. Paris Ser. I Math., 322 (1996), pp. 497–502.

[12] J. Jarusek and Ch. Eck, *Dynamic contact problems with small Coulomb friction for viscoelastic bodies. Existence of solutions*, Math. Models Methods Appl. Sci., 9 (1999), pp. 11–34.

[13] A. Klarbring, A. Mikelic, and M. Shillor, *Frictional contact problems with normal compliance*, Internat. J. Engrg. Sci., 26 (1988), pp. 811–832.

[14] N. Kikuchi and T. J. Oden, *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM Stud. Appl. Math. 8, Philadelphia, 1988.

[15] K. L. Kuttler, *Dynamic friction contact problems for general normal and friction laws*, Nonlinear Anal., 28 (1997), pp. 559–575.

[16] K. L. Kuttler, *Modern Analysis*, CRC Press, Boca Raton, FL, 1998.

[17] K. L. Kuttler, Y. Renard, and M. Shillor, *Models and simulations of dynamic frictional contact of a beam*, Comput. Methods Appl. Mech. Engrg., 177 (1999), pp. 259–272.

[18] K. L. Kuttler and M. Shillor, *Set-valued pseudomonotone maps and degenerate evolution inclusions*, Commun. Contemp. Math., 1 (1999), pp. 87–123.

[19] K. L. Kuttler and M. Shillor, *Dynamic bilateral contact with discontinuous friction coefficient*, Nonlinear Anal., 45 (2001), pp. 309–327.

[20] J. L. Lions, *Quelques Methods de Resolution des Problemes aux Limites Non Lineaires*, Dunod, Paris, 1969.

[21] J. A. C. Martins and J. T. Oden, *Existence and uniqueness results for dynamic contact problems with nonlinear normal and friction interface laws*, Nonlinear Anal., 11 (1987), pp. 407–428.

[22] Z. Naniewicz and P. D. Panagiotopoulos, *Mathematical Theory of Hemivariational Inequalities and Applications*, Marcel Dekker, New York, 1995.

[23] O. A. Oleinik, A. S. Shamaev, and G. A. Yosifian, *Mathematical Problems in Elasticity and Homogenization*, North-Holland, Amsterdam, 1992.

[24] P. D. Panagiotopoulos, *Inequality Problems in Mechanics and Applications*, Birkhäuser, Basel, 1985.

[25] E. Rabinowiz, *Friction and Wear of Materials*, 2nd ed., Wiley, New York, 1995.

[26] M. Rochdi and M. Shillor, *A Dynamic Thermoviscoelastic Frictional Contact Problem with Damped Response*, preprint, 1998.

[27] M. Rochdi, M. Shillor, and M. Sofonea, *Quasistatic viscoelastic contact with normal compliance and friction*, J. Elasticity, 51 (1998), pp. 105–126.

[28] T. I. Seidman, *The transient semiconductor problem with generation terms*, II, in Nonlinear Semigroups, Partial Differential Equations and Attractors, Lecture Notes in Math. 1394, Springer-Verlag, New York, 1989, pp. 185–198.

[29] M. Shillor, ed., *Recent advances in contact mechanics*, Math. Comput. Modelling, 28 (1998).

[30] M. Sofonea and M. Shillor, *Quasistatic viscoelastic contact with friction*, Internat. J. Engrg. Sci., 38 (2000), pp. 1517–1533.

[31] N. Strömberg, L. Johansson, and A. Klarbring, *Derivation and analysis of a generalized standard model for contact friction and wear*, Internat. J. Solids Structures, 33 (1996), pp. 1817–1836.

[32] W. R. D. Wilson, *Modeling friction in sheet-metal forming simulation*, in The Integration of Materials, Processes and Product Design, Zabaras et al., eds., Balkema, Rotterdam, 1999, pp. 139–147.

SIAM J. MATH. ANAL.
Vol. 34, No. 1, pp. 28–63

UNIQUENESS AND STABILITY OF L^∞ SOLUTIONS FOR TEMPLE CLASS SYSTEMS WITH BOUNDARY AND PROPERTIES OF THE ATTAINABLE SETS*

FABIO ANCONA† AND PAOLA GOATIN‡

Abstract. We consider the initial-boundary value problem for a strictly hyperbolic, genuinely nonlinear, Temple class system of conservation laws

$$u_t + f(u)_x = 0, \qquad u \in \mathbb{R}^n,$$

on the domain $\Omega = \{(t,x) \in \mathbb{R}^2 : t \geq 0,\, x \geq 0\}$. For a class of initial data $\overline{u} \in \mathbf{L}^\infty(\mathbb{R}^+)$ and boundary data $\widetilde{u} \in \mathbf{L}^\infty(\mathbb{R}^+)$ with possibly unbounded variation, we construct a flow of solutions $(\overline{u}, \widetilde{u}) \to u(t) \doteq E_t(\overline{u}, \widetilde{u})$ that depend continuously, in the $\mathbf{L}^1$ distance, both on the initial data and on the boundary data. Moreover, we show that each trajectory $t \mapsto E_t(\overline{u}, \widetilde{u})$ provides the unique weak solution of the corresponding initial-boundary value problem that satisfies an entropy condition of Oleinik type.

Next, we study the initial-boundary value problem for the above equation from the point of view of control theory taking the initial data $\overline{u}$ fixed and considering, in connection with a prescribed set $\mathcal{U}$ of boundary data regarded as admissible controls, the set of attainable profiles at a fixed time $T > 0$, and at a fixed point $\overline{x} > 0$:

$$\mathcal{A}(T, \mathcal{U}) \doteq \left\{E_T(\overline{u}, \widetilde{u})(\cdot)\ ;\ \widetilde{u} \in \mathcal{U}\right\}, \qquad \mathcal{A}(\overline{x}, \mathcal{U}) \doteq \left\{E_{(\cdot)}(\overline{u}, \widetilde{u})(\overline{x})\ ;\ \widetilde{u} \in \mathcal{U}\right\}.$$

We establish closure and compactness of the sets $\mathcal{A}(T, \mathcal{U})$, $\mathcal{A}(\overline{x}, \mathcal{U})$ in the $\mathbf{L}^1_{loc}$ topology for a class $\mathcal{U}$ of admissible controls satisfying convex constraints.

Key words. hyperbolic systems, conservation laws, Temple class systems, Lipschitz semigroup, boundary control, attainable set

AMS subject classifications. 35L65, 35B37

PII. S0036141001383424

1. Introduction.

Consider the initial-boundary value problem for a nonlinear, strictly hyperbolic system of conservation laws in one space dimension,

$$u_t + f(u)_x = 0, \tag{1.1}$$

$$u(0,x) = \overline{u}(x), \tag{1.2}$$

$$u(t,0) = \widetilde{u}(t), \tag{1.3}$$

on the domain $\Omega = \{(t,x) \in \mathbb{R}^2\ ;\ t \geq 0,\, x \geq 0\}$. Here, $u = u(t,x) \in \mathbb{R}^n$ is the vector of the conserved quantities, and the flux function $f : U \mapsto \mathbb{R}^n$ is a smooth vector field defined on some open set $U \subseteq \mathbb{R}^n$. We recall that, for problems of this type, classical solutions may develop discontinuities in finite time, no matter the regularity of the initial and boundary data. Hence, it is natural to consider weak solutions in the sense of distributions. Moreover, in general, the Dirichlet condition (1.3) cannot be fulfilled pointwise a.e. (see [7, 17]), even when (1.1) is a linear system (cf. [23]).

*Received by the editors January 9, 2001; accepted for publication (in revised form) March 25, 2002; published electronically August 15, 2002. This research was partially supported by the European TMR Network on Hyperbolic Conservation Laws ERBFMRXCT960033.

http://www.siam.org/journals/sima/34-1/38342.html

†Dipartimento di Matematica and C.I.R.A.M., Piazza Porta S. Donato, n. 5, 40123 Bologna, Italy (ancona@ciram3.ing.unibo.it).

‡Centre de Mathématiques Appliquées and Centre National de la Recherche Scientifique, U.M.R. 7641, Ecole Polytechnique, 91128 Palaiseau Cedex, France (goatin@cmapx.polytechnique.fr).

For this reason, different weaker formulations of the boundary condition have been considered in the literature, both for characteristic boundaries (where the eigenvalues of the Jacobian matrix $Df(u)$ may coincide with the slopes of the boundary profile) and for noncharacteristic ones; see [1, 28] and the references therein.

In this paper we assume that the boundary is noncharacteristic, requiring that, for each ith characteristic family, the ith eigenvalue $\lambda_i(u)$ of $Df(u)$ always has the same sign, which implies that there is a fixed set of characteristic lines entering the interior of the domain Ω at any point of the boundary $x = 0$. Then, following Dubois and LeFloch [17] and Joseph and LeFloch [21], we reformulate (1.3) in the weak form

$$f\big(u(t,0+)\big) \in f\big(\mathcal{V}(\widetilde{u}(t))\big), \qquad \text{for a.e. } t>0, \tag{1.4}$$

where $\mathcal{V}(\widetilde{u}(t)) \subset U$ is a time-dependent set (the set of *admissible boundary values*) that is defined from the boundary data $\widetilde{u}$ using the notion of Riemann problem, while $f(u(t,0+))$ represents the weak trace of the flux $f(u(t,x))$ at $x = 0$. We are concerned with the well-posedness of (1.1)–(1.2), (1.4) within domains of $\mathbf{L}^\infty$ functions with possibly unbounded variations, having in mind to study the initial-boundary value problem (1.1)–(1.2), (1.4) from the point of view of control theory where it is natural to regard the boundary data as varying within a prescribed set of admissible $\mathbf{L}^\infty$ controls.

We recall that, for initial and boundary data with small total variation, the existence of global weak solutions of the corresponding mixed problem for (1.1), with various types of boundary conditions, was studied by Liu [25, 26], Goodman [19], Dubroca and Gallice [18], and Sablé-Tougeron [28], using the Glimm scheme, and by Amadori [1], developing a front tracking algorithm. More recently, the Lipschitz continuous dependence on the initial and boundary data of entropy admissible BV solutions was obtained in [2, 3], for systems of two equations, following the *semigroup approach* developed by Bressan and his collaborators to prove the well-posedness of the Cauchy problem for (1.1) (see [11]).

Notice that, while for scalar conservation laws the well-posedness theory for the mixed problem had been established within domains of $\mathbf{L}^\infty$ functions [23, 24, 29], in the case of systems the available stability results apply only to solutions with small total variation. In the present paper we extend these results to domains of $\mathbf{L}^\infty$ functions for a class of systems introduced by Temple [30, 29] in which rarefaction and Hugoniot curves coincide, under the assumption that all characteristic fields are genuinely nonlinear in the sense of Lax. Namely, for such systems, and for a domain $\mathcal{D}$ of pairs of $\mathbf{L}^\infty$ functions with possibly unbounded variation, we construct a continuous flow of solutions

$$(\,\overline{u},\,\widetilde{u}\,) \mapsto u(t,\,\cdot\,) \doteq E_t(\,\overline{u},\,\widetilde{u}\,)(\cdot), \qquad (\,\overline{u},\,\widetilde{u}\,) \in \mathcal{D} \tag{1.5}$$

of the mixed problem (1.1)–(1.2), (1.4), that, for every fixed $\delta > 0$, satisfy the stability estimate

$$\begin{aligned} &\big\|E_t(\,\overline{u},\,\widetilde{u}\,) - E_t(\,\overline{v},\,\widetilde{v}\,)\big\|_{\mathbf{L}^1([\delta,+\infty[)} \\ &\qquad \leq L_t \cdot \Big\{ \big\|\,\overline{u}-\overline{v}\,\big\|_{\mathbf{L}^1(\mathbb{R}^+)} + \big\|f(\widetilde{u})-f(\widetilde{v})\big\|_{\mathbf{L}^1([0,\,t])} \Big\} \end{aligned} \tag{1.6}$$

for all $t \geq \delta$, where the Lipschitz constant L_t takes the form $L_t = C(1+\log(t/\delta))$, for some constant $C > 0$ depending on the system (1.1). Moreover, relying on a

stability estimate of the same type, established for the map $(\,\overline{u},\,\widetilde{u}\,)\mapsto E_{(\cdot)}(\,\overline{u},\,\widetilde{u}\,)(x)$, $x>0$, we prove that every solution $u(t,x)\doteq E_t(\,\overline{u},\,\widetilde{u}\,)(x)$ actually admits a strong $\mathbf{L}^1$ trace at the boundary $x=0$. In the same spirit of [6, 14], the flow map E_t in (1.5) is constructed as the unique limit of a Cauchy sequence of flow maps E^ν_t whose trajectories are front tracking approximate solutions of (1.1) in the region Ω that satisfy a stability estimate of the same type as (1.6) (with a Lipschitz constant independent of ν and on the total variation).

Concerning the existence of weak solutions of the initial-boundary value problem (1.1)–(1.3) for Temple class systems with $\mathbf{L}^\infty$ data, an earlier result can be found in [16], where it was shown the convergence of the viscous approximate solutions using a compensated compactness argument.

In order to obtain the well-posedness of the mixed problem (1.1)–(1.2), (1.4) with initial and boundary data $(\,\overline{u},\,\widetilde{u}\,)\in\mathcal{D}$, we next show that a distributional solution $u=u(t,x)$ of (1.1)–(1.2), (1.4) coincides with the corresponding trajectory of the flow map E_t if and only if, letting $w=(w_1,\ldots,w_n)$ denote a system of Riemann coordinates for (1.1), and assuming that the characteristic speeds entering the domain Ω are λ_i, $i\in\{n-p+1,\ldots,n\}$, the following conditions hold:

(i) The map $(t,x)\to\big(u(t,\cdot\,),\,u(\cdot\,,x)\big)$ takes values within the domain $\mathcal{D}$.

(ii) u satisfies suitable Oleinik-type conditions on the decay of positive waves in time and in space.

(iii) u admits at $t=0$ and at $x=0$ the essential limits

$$\operatorname*{ess\ sup}_{t\to 0^+}\big\|u(t,\cdot\,)-\overline{u}\,\big\|_{\mathbf{L}^1([0,R])}=0\,,$$

$$\operatorname*{ess\ sup}_{x\to 0^+}\big\|w_i\big(u(\cdot\,,x)\big)-w_i\circ\widetilde{u}\big\|_{\mathbf{L}^1([0,\,\tau])}=0\qquad\forall\ i\in\{n-p+1,\ldots,n\}$$

for any $R>0$, $\tau>0$.

Relying on the formulation of the boundary condition in terms of *boundary entropy pairs*, introduced by Otto [27] for scalar conservation laws, and then extended by Chen and Frid [15, 16] to various classes of systems (including Temple systems), one can recover the regularity conditions (iii) employing the corresponding distributional entropy inequality. We thus prove, in particular, that any weak solution of the mixed problem (1.1)–(1.2), (1.4) constructed by the Glimm scheme or by a front tracking algorithm, which clearly satisfies the Oleinik-type conditions (ii) and any entropy inequality, must coincide with the trajectory of the flow map E_t in (1.5).

The proof of the $\mathbf{L}^1$ stability estimate (1.6) is based on the same homotopy and linearization technique developed in [12, 6, 14]. In order to estimate how the distance between two infinitesimally close solutions varies in time, the basic idea consists of "differentiating" a family of front tracking approximate solutions w.r.t. a parameter which determines the (space) locations of the jumps and in providing a priori bounds on the norm of the resulting "shift differential". In particular, given a piecewise constant solution $u=u(t,x)$ with an initial data $\overline{u}$ and a boundary data $\widetilde{u}$, we may consider a family of perturbed solutions $\theta\mapsto u^\theta(t,\cdot\,)$ obtained from u by shifting the time position t_0 of a single jump in $\widetilde{u}$ at constant rate $\widetilde{\xi}$ (or the space location x_0 of a single jump in $\overline{u}$ at constant rate ξ_0). Call $\widetilde{\sigma}\doteq f(\widetilde{u}(t_0+))-f(\widetilde{u}(t_0-))$ the size of this jump in $f(\widetilde{u})$ (or let $\sigma\doteq\overline{u}(x_0+)-\overline{u}(x_0-)$ denote the size of the jump in $\overline{u}$). As long as the wave-front configuration of the perturbed and unperturbed solutions is the same, for every fixed $\delta>0$, and any $t\geq\delta$, one can estimate the $\mathbf{L}^1$ distance between $u(t,\cdot\,)\restriction_{[\delta,+\infty[}$ and $u^\theta(t,\cdot\,)\restriction_{[\delta,+\infty[}$ by showing that, if the perturbed solution

$u^\theta(t,\cdot)$ contains jumps of size $\sigma_1^\theta,\dots,\sigma_M^\theta$ located at $x_\alpha^\theta \geq \delta$, and shifted at shift rate (in space) $\xi_1^\theta,\dots,\xi_M^\theta$, then there holds

$$\sum_{\alpha=1}^{M} \left|\sigma_\alpha^\theta\, \xi_\alpha^\theta\right| \leq L_t \cdot \left|\widetilde{\sigma}\, \widetilde{\xi}\right| \tag{1.7}$$

for some constant $L_t = L_t(\delta) > 0$ depending on t and on the distance δ from the boundary $x = 0$. The key stability estimate (1.7) is obtained here as in [14], relying on two remarkable properties of genuinely nonlinear systems of Temple class:

(a) By genuine nonlinearity and finite propagation speed, for every fixed $\delta > 0$, the total amount of waves in a solution $u(t,\cdot)$ to the mixed problem (1.1)–(1.2), (1.4), which can be influenced by shifting a single wave-front of the initial data $\overline{u}$ or of the boundary data $\widetilde{u}$, and are located at distance $\geq \delta$ from the boundary $x = 0$, remains uniformly bounded, for all $t \geq \delta$, by some constant depending on t and δ.

(b) For solutions of Temple class systems, the support of perturbations satisfies a special localization property related to the representation formula for the solutions of the nonlinear equation $U_t + f(U_x) = 0$ in terms of envelopes of n families of hyperplanes [29].

Having in mind applications of Temple systems to problems of oil reservoir simulation, multicomponent chromatography, as well as in models for traffic flows, in the last part of the paper we focus our attention on the mixed problem (1.1)–(1.2), (1.4) from the point of view of control theory. Namely, following the same approach adopted by Ancona and Marson [4, 5] for scalar conservation laws, we fix an initial data $\overline{u} \in \mathbf{L}^\infty(\mathbb{R}^+)$, and, in connection with an assigned set $\mathcal{U} \subset \mathbf{L}^\infty(\mathbb{R}^+)$ of boundary data regarded as admissible controls, we consider the sets of attainable profiles at a fixed time T,

$$\mathcal{A}(T,\mathcal{U}) \doteq \big\{E_T(\,\overline{u},\,\widetilde{u}\,)(\cdot)\ ;\ \widetilde{u} \in \mathcal{U}\big\},$$

and at a fixed point in space $\overline{x} > 0$,

$$\mathcal{A}(\,\overline{x},\mathcal{U}) \doteq \big\{E_{(\cdot)}(\,\overline{u},\,\widetilde{u}\,)(\overline{x})\ ;\ \widetilde{u} \in \mathcal{U}\big\}\,.$$

Relying on the well-posedness theory provided by the previous results, we establish here, as in the scalar case [4, 5], the compactness of these sets in the $\mathbf{L}^1{}_{loc}$ topology for a class $\mathcal{U}$ of admissible boundary controls that satisfy convex constraints.

The paper is organized as follows. Section 2 contains the basic definitions and the statement of the main results. In section 3 we provide an outline of the proof of Theorem 2.3 on the existence of a continuous flow of solutions depending continuously, in the $\mathbf{L}^1$ distance, on the initial and on the boundary data. The basic a priori estimates on shift differentials are contained in sections 4 and 5, while the proof of Theorem 2.3 is given in section 6, which also contains the proof of Theorem 2.4. The compactness property of the attainable sets stated in Theorem 2.6 is established in section 7.

2. Preliminaries and statement of the main results.

2.1. Formulation of the problem. Let $f : U \mapsto \mathbb{R}^n$ be the flux function of the strictly hyperbolic system (1.1) defined on a neighborhood of the origin $U \subseteq \mathbb{R}^n$, and denote by $\lambda_1(u) < \cdots < \lambda_n(u)$ the eigenvalues of the Jacobian matrix $Df(u)$.

Choose right and left eigenvectors $r_i(u)$, $l_i(u)$, $i = 1, \dots, n$, of $Df(u)$ normalized so that

$$|r_i(u)| = 1, \qquad l_i(u) \cdot r_j(u) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{2.1}$$

By possibly considering a sufficiently small restriction of the domain U, we may assume that a *uniform* strict hyperbolicity condition holds:

(SH1) For every $u, v \in U$, the characteristic speeds at these points satisfy

$$\lambda_j(u) > \lambda_i(v) \qquad \forall\, 1 \leq i < j \leq n\,. \tag{2.2}$$

We also assume that each ith characteristic field r_i is *genuinely nonlinear* in the sense of Lax, i.e., that, by choosing a suitable orientation of the eigenvectors $r_i(u)$, at every point $u \in U$ one has $D\lambda_i \cdot r_i(u) > 0$. Moreover, the system (1.1) is of Temple class in accordance with the following.

Definition 2.1. *A system of conservation laws is of Temple class if there exists a system of coordinates $w = (w_1, \dots, w_n)$ consisting of Riemann invariants and such that the level sets $\{u \in U;\ w_i(u) = \text{constant}\}$ are hyperplanes (see* [29]*).*

It is not restrictive to assume that the Riemann coordinates are chosen so that $(w_1, \dots, w_n)(0) = (0, \dots, 0)$ and

$$\frac{\partial}{\partial w_i}\lambda_i(w) > 0 \qquad \forall\, w = w(u), \quad u \in U, \qquad i = 1, \dots, n. \tag{2.3}$$

Throughout the paper, we will often write $w_i(t,x) \doteq w_i\big(u(t,x)\big)$ to denote the ith Riemann coordinate of a solution $u = u(t,x)$ to (1.1). For a Temple class system, the integral curve of the vector field r_i through a point u_0 is the straight line described by the $n-1$ linear equations

$$w_j(u) = w_j(u_0)\,, \qquad j \neq i. \tag{2.4}$$

In particular, shock and rarefaction curves coincide. Let $\sigma \mapsto R_i(u)[\sigma]$ denote the ith rarefaction curve through $u \in U$. We fix a convex, compact set $K \subset U$ having the form

$$K = \Big\{u \in U; \quad w_i(u) \in [a_i,\ b_i], \qquad i = 1, \dots, n\Big\}, \tag{2.5}$$

and, concerning the boundary, we assume that there is a fixed set of characteristic lines entering the interior of the domain Ω at every point of the boundary $x = 0$, i.e., that, for some index $p \in \{1, \dots, n\}$, there holds

$$\lambda_{n-p}(u) < 0 < \lambda_{n-p+1}(u) \qquad \forall\, u \in K. \tag{2.6}$$

We shall denote by $\lambda^{\min}$, $\lambda^{\max}$ the minimum and maximum characteristic speed so that there holds

$$0 < \lambda^{\min} \leq |\lambda_i(u)| \leq \lambda^{\max} \qquad \forall\, u \in K, \qquad \forall\, i \in \{1, \dots, n\}. \tag{2.7}$$

Remark 2.1. Since the rarefaction curves are straight lines, the existence of Riemann coordinates implies

$$r_k\big(R_i(u)[\sigma]\big) \in \operatorname{span}\big\{r_k(u),\, r_i(u)\big\} \qquad \forall\, u \in U, \quad \forall\, \sigma, \quad \forall\, i, k = 1, \dots, n\,. \tag{2.8}$$

Relying on this property, one can easily verify that a strengthened version of the strict hyperbolicity assumption on the linear independence of the eigenvectors $\{r_1(u), \dots, r_n(u)\}$, $u \in U$, holds, namely:

(SH2) Given any n-tuple of states $u^1, \dots, u^n \in U$, such that

$$u^{i+1} = R_i(u^i)[\sigma_i], \qquad 1 \leq i < n\,,$$

for some $\sigma_1, \dots, \sigma_n$, the eigenvectors $r_1(u^1), \dots, r_n(u^n)$ are linearly independent.

Notice that the strict hyperbolicity condition (SH2) implies the invertibility of the map $f : U \mapsto f(U)$. Indeed, for any given pair of states $u, v \in U$, there will be some values $\sigma_1, \dots, \sigma_n$ so that, setting

$$z^1 \doteq u, \qquad z^{i+1} \doteq z^i + \sigma_i\, r_i(z^i), \qquad 1 \leq i \leq n,$$

we can write

$$v - u = \sum_{i=1}^n \sigma_i\, r_i(z^i). \tag{2.9}$$

In turn, (2.9) yields

$$f(v) - f(u) = \sum_{i=1}^n \sigma_i\, \lambda_i(z^i,\, z^{i+1})\, r_i(z^i), \tag{2.10}$$

where $\lambda_i(z^i,\, z^{i+1})$ denotes the ith eigenvalue of the averaged matrix

$$A(z^i,\, z^{i+1}) = \int_0^{\sigma_i} Df(z^i + \theta r_i(z^i))\; d\theta\,. \tag{2.11}$$

Observing that, by the genuine nonlinearity of the characteristic speeds and because of the assumption (2.6), one has

$$\sigma_i \neq 0 \quad \Longrightarrow \quad \lambda_i(z^i,\, z^{i+1}) \neq 0 \qquad \forall\; i = 1, \dots, n\,,$$

using (SH2) one clearly deduces from (2.9)–(2.10) the injectivity of the flux function f.

We next introduce a definition of weak solution to (1.1)–(1.3) which includes an entropy admissibility condition of Oleinik type on the decay of positive waves. The boundary condition is formulated in terms of the weak trace of $f(u)$ at the boundary $x = 0$ and is related to the notion of Riemann problem in the same spirit of [17]. To this purpose, letting $u(t,x) = W(\xi = x/t;\, u_L, u_R)$, $u_L, u_R \in K$, denote the self-similar solution of the Riemann problem for (1.1) with initial data

$$u(0,x) = \begin{cases} u_L & \text{if } x < 0, \\ u_R & \text{if } x > 0, \end{cases}$$

for any given boundary state $\widetilde{u} \in K$, we define the set of *admissible states at the boundary*

$$\mathcal{V}(\widetilde{u}) := \big\{ W(0+;\, \widetilde{u}, u_R)\;;\; u_R \in K \big\}. \tag{2.12}$$

Definition 2.2. *A function $u : [0,\, T] \times \mathbb{R}^+ \mapsto K$ is an entropy weak solution of the initial-boundary value problem* (1.1)–(1.3) *on $\Omega_T \doteq [0,\, T] \times \mathbb{R}^+$ if it is continuous as a function from $]0,\, T]$ into $\mathbf{L}^1{}_{loc}$ and the following properties hold:*

(i) *u is a distributional solution to the Cauchy problem* (1.1)–(1.2) *on Ω_T in the sense that, for every test function $\phi \in \mathcal{C}^1_c$ with compact support contained in the set $\{(t,x) \in \mathbb{R}^2;\ x > 0,\, t < T\}$, there holds*

$$\int_0^T \int_0^{+\infty} \big(u(t,x)\cdot\phi_t(t,x) + f(u(t,x))\cdot\phi_x(t,x)\big)dx\,dt + \int_0^{+\infty} \overline{u}(x)\cdot\phi(0,x)dx = 0.$$

(ii) *The flux $f(u)$ admits a weak* trace at the boundary $x = 0$; i.e., there exists a measurable function $\Psi : [0,\, T] \mapsto \mathbb{R}^n$ such that*

$$f(u(\cdot,x)) \ \underset{x\to 0^+}{\overset{*}{\longrightarrow}}\ \Psi \qquad \text{in} \quad \mathbf{L}^\infty([0,\,T]), \tag{2.13}$$

and the boundary condition (1.3) *is satisfied in the following sense:*

$$\Psi(t) \in f\big(\mathcal{V}(\widetilde{u}(t))\big) \qquad \text{for a.e.} \quad 0 \le t \le T. \tag{2.14}$$

(iii) *u satisfies the following entropy conditions on the decay of positive waves in time and in space. There exists some constant $C > 0$, depending only on the system* (1.1), *so that*

(a) *for any $0 < t \le T$, and for a.e. $0 < x < y$, there holds*

$$w_i(t,y) - w_i(t,x) \ \le\ C \cdot \frac{y-x}{t} \qquad \text{if} \qquad i \in \{1,\dots,n-p\}, \tag{2.15}$$

$$w_i(t,y) - w_i(t,x) \ \le\ C \cdot \left\{ \frac{y-x}{t} + \log\frac{y}{x} \right\} \text{if } i \in \{n-p+1,\dots,n\}; \tag{2.16}$$

(b) *for a.e. $x > 0$, and for a.e. $0 < \tau_1 < \tau_2 \le T$, there holds*

$$w_i(\tau_2,x) - w_i(\tau_1,x) \ \le\ C \cdot \log\frac{\tau_2}{\tau_1} \qquad \text{if} \qquad i \in \{1,\dots,n-p\}, \tag{2.17}$$

$$w_i(\tau_2,x) - w_i(\tau_1,x) \ \le\ C \cdot \left\{ \frac{\tau_2-\tau_1}{x} + \log\frac{\tau_2}{\tau_1} \right\} \text{if } i \in \{n-p+1,\dots,n\}. \tag{2.18}$$

Remark 2.2. The set of admissible flux values at the boundary $f\big(\mathcal{V}(\widetilde{u})\big)$ can be expressed in Riemann coordinates as

$$f\big(\mathcal{V}(\widetilde{u})\big) = \Big\{ f(u)\ ;\ w_j(u) = w_j(\widetilde{u}) \qquad \forall\ j = n-p+1,\dots,n \Big\}. \tag{2.19}$$

Hence, by the invertibility of the map $f : U \mapsto f(U)$, the above boundary condition (2.14) is equivalent to the set of equalities

$$w_j\big(f^{-1}(\Psi(t))\big) = w_j\big(\widetilde{u}(t)\big) \quad \text{ for a.e.} \quad 0 \le t \le T, \quad j = n-p+1,\dots,n. \tag{2.20}$$

This means that the boundary condition (2.14) guarantees that, at almost every time $t \in [0,\, T]$, the solution to the Riemann problem for (1.1), having left and right

initial states $u^L = \widetilde{u}(t)$, $u^R = f^{-1}(\Psi(t))$, contains only waves with negative speeds and, in particular, its restriction to the region $[t, +\infty]\times]0, +\infty[$ takes constant value $f^{-1}(\Psi(t))$.

Remark 2.3. Several definitions of sets of admissible boundary values that can be used for alternative formulations of the boundary condition (2.14) have been proposed in the literature. (A systematic study of such formulations is contained in [21].) In particular, following Dubois and LeFloch [17] one may consider an admissible set $\mathcal{V}^{\mathcal{E}ntr}$ whose definition is based on the boundary entropy inequalities associated with the artificial vanishing viscosity limit

$$u^\varepsilon_t + f(u^\varepsilon)_x = \varepsilon u^\varepsilon_{xx}, \qquad \varepsilon \to 0\,. \tag{2.21}$$

Namely, in accordance with [17], for any given boundary state $\widetilde{u}$ the set of admissible boundary values $\mathcal{V}^{\mathcal{E}ntr}(\widetilde{u})$ based on the vanishing viscosity limit (2.21) is defined as

$$\begin{aligned}\mathcal{V}^{\mathcal{E}ntr}(\widetilde{u}) \doteq \Big\{ u \ ;\ &\text{for all convex entropy-entropy flux pairs } (\eta, q)\\ &q(u) - q(\widetilde{u}) - D\eta(\widetilde{u})\big(f(u) - f(\widetilde{u})\big) \le 0 \Big\}\,.\end{aligned} \tag{2.22}$$

The resulting boundary condition (2.14), with $f(\mathcal{V}(\widetilde{u}))$ replaced by $f(\mathcal{V}^{\mathcal{E}ntr}(\widetilde{u}))$, is a generalization of the earlier one introduced by Bardos, Leroux, and Nedelec in [7] for scalar (multidimensional) conservation laws, which used only the boundary entropy inequalities associated with the Kruzkov entropies to define the set (2.22). By reformulating the entropy inequalities in terms of Young measures (associated with a sequence of viscous approximates solutions) it is shown in [21] that the boundary condition (2.14) corresponding to the set of admissible boundary data (2.22) is satisfied by any (artificial) vanishing viscosity limit (2.21). Thus, this formulation of the boundary condition is natural at least in the case that no boundary layer develops near the boundary. Moreover, it is proved in [17, 21] that, for linear hyperbolic systems and scalar conservation laws, the two sets of admissible boundary data (2.12), (2.22) are the same, and hence the two formulations of the boundary condition are equivalent. Indeed, as it was conjectured by Dubois and LeFloch [17], the two sets (2.12), (2.22) also coincide in the case of Temple systems. The inclusion $\mathcal{V}(\widetilde{u}) \subseteq \mathcal{V}^{\mathcal{E}ntr}(\widetilde{u})$ was proved by Benabdallh and Serre [8], while the converse inclusion $\mathcal{V}^{\mathcal{E}ntr}(\widetilde{u}) \subseteq \mathcal{V}(\widetilde{u})$ can be established making use of the Kruzkov-type entropies associated with a Temple system, as it is shown in the appendix.

An alternative formulation of the boundary condition (1.3) for $\mathbf{L}^\infty$ boundary data, which is not based on the existence of the trace of the solution (or of the flux of the solution) at the boundary, was proposed by Otto [27], for scalar conservation laws, following the vanishing viscosity approach, and then extended by Chen and Frid [15, 16] to various classes of systems (including Temple systems). In this case, the boundary condition is expressed, requiring that the solution satisfy a family of boundary entropy admissibility integral inequalities that are associated with the *boundary entropy pairs* for the system (1.1). Applying the theory of divergence measure fields, it is shown in [15] that, for scalar conservation laws and for Temple class sytems, a solution satisfying such an integral formulation of the boundary conditions assumes the boundary data also in the sense of our Definition 2.2.

2.2. Stability and uniqueness of weak solutions. Due to the presence of the boundary data, the flow map $u(0,\cdot) \mapsto u(t,\cdot)$ induced by (1.1)–(1.3) is not time homogeneous. To recast the problem in a semigroup framework, it is thus convenient

to incorporate the boundary data $\widetilde{u}$ in the domain of the semigroup. More precisely, in connection with a convex, compact set $K \subset U$ of the form (2.5), we consider the positively invariant domain of pairs of $\mathbf{L}^\infty$ functions, with possibly unbounded variations

$$\mathcal{D} \doteq \Big\{ \mathbf{p} = (\,\overline{u},\, \widetilde{u}\,)\ ;\ \overline{u},\, \widetilde{u} \in \mathbf{L}^1(\mathbb{R}^+, K) \Big\}, \tag{2.23}$$

where $\mathbf{L}^1(\mathbb{R}^+, K)$ denotes the metric space of all $\mathbf{L}^1$ functions $u : \mathbb{R}^+ \mapsto K$, equipped with the usual $\mathbf{L}^1$ distance. Let $\mathcal{T}_t : \mathbf{L}^1(\mathbb{R}^+, K) \mapsto \mathbf{L}^1(\mathbb{R}^+, K)$ be the translation operator, i.e., $(\mathcal{T}_t\widetilde{u})(s) \doteq \widetilde{u}(t+s)$, and denote by $E : \mathbb{R}^+ \times \mathcal{D} \mapsto \mathbf{L}^1(\mathbb{R}^+, K)$ the evolution operator $E_t\mathbf{p} = u(t,\cdot)$, u being a solution to (1.1)–(1.3). With the above notations, we shall construct a semigroup S acting on $\mathcal{D}$, in the sense that

$$\begin{array}{rcl} S : \mathbb{R}^+ \times \mathcal{D} & \mapsto & \mathcal{D}, \\ (\,t\,,\,\mathbf{p}\,) & \mapsto & S_t\mathbf{p}, \end{array} \tag{2.24}$$

where, if $\mathbf{p} = (\,\overline{u},\, \widetilde{u}\,)$, $S_t\mathbf{p} = \big(E_t\mathbf{p},\, \mathcal{T}_t\widetilde{u}\,\big)$.

Our main result is concerned with the existence of an $\mathbf{L}^1$ continuous semigroup of the form (2.24), generated by the system (1.1) on the domain $\mathcal{D}$.

THEOREM 2.3. *Let* (1.1) *be a system of Temple class with all characteristic fields genuinely nonlinear. Assume that* (2.6) *and the strict hyperbolicity condition* (SH1) *are verified. Then there exist a continuous semigroup S of the form* (2.24) *and some constant $C > 0$, depending only on the system* (1.1) *and on the domain K, so that, for every fixed $\delta > 0$, and for all $(\,\overline{u},\, \widetilde{u}\,), (\,\overline{v},\, \widetilde{v}\,) \in \mathcal{D}$, letting $L_t \doteq L_t(\delta) = C(1+\log(t/\delta))$, one has*

$$\begin{aligned} &\big\| E_t(\,\overline{u},\, \widetilde{u}\,) - E_t(\,\overline{v},\, \widetilde{v}\,) \big\|_{\mathbf{L}^1([\delta, +\infty[)} \\ &\qquad \leq L_t \cdot \Big\{ \big\| \,\overline{u} - \overline{v}\, \big\|_{\mathbf{L}^1(\mathbb{R}^+)} + \big\| f(\widetilde{u}) - f(\widetilde{v}) \big\|_{\mathbf{L}^1([0,\,t])} \Big\} \end{aligned} \tag{2.25}$$

for all $t \geq \delta$. Moreover, the map $(t,x) \mapsto E_t(\,\overline{u},\, \widetilde{u}\,)(x)$ yields an entropy weak solution (in the sense of Definition 2.2*) to the initial-boundary value problem* (1.1)–(1.3) *on Ω that admits a strong $\mathbf{L}^1$ trace at the boundary $x = 0$; i.e., there exists a measurable map $\psi : \mathbb{R}^+ \mapsto U$ such that*

$$\lim_{x \to 0^+} \int_0^\tau \big| E_t(\,\overline{u},\, \widetilde{u}\,)(x) - \psi(t) \big|\, dt = 0 \qquad \forall\ \tau \geq 0. \tag{2.26}$$

Remark 2.4. With the same arguments used in section 6 to establish the existence of a strong $\mathbf{L}^1$ trace at the boundary $x = 0$ for the map $(t,x) \mapsto E_t(\,\overline{u},\, \widetilde{u}\,)(x)$ provided by Theorem 2.3, one can show the continuity w.r.t. the $\mathbf{L}^1{}_{loc}$ topology of $t \mapsto E_t(\,\overline{u},\, \widetilde{u}\,)$ at time $t = 0$. However, as in the case of the Cauchy problem [14], the map $t \mapsto E_t(\,\overline{u},\, \widetilde{u}\,)$ may not be Lipschitz continuous at $t = 0$ if the initial condition $\overline{u}$ has unbounded total variation. Moreover, the evolution operator $\mathbf{p} \mapsto E_t\mathbf{p}$ is not, in general, Lipschitz continuous w.r.t. the topology of $\mathbf{L}^1(\mathbb{R}^+, K)$ on the range $E_t(\mathcal{D})$.

Remark 2.5. The trajectories of the flow map E_t provided by Theorem 2.3 are obtained as limits of front tracking approximate solutions whose values are independent of the Riemann coordinates of the boundary data that leave the domain Ω. Thus, the same property holds for the limit map E_t, and hence, given any couple of initial data

and boundary condition $(\,\overline{u},\,\widetilde{u}\,)\in\mathcal{D}$, if we consider the auxiliary boundary condition $\widetilde{u}'$ defined in Riemann coordinates by

$$w_j(\widetilde{u}'(t))\doteq\begin{cases}\overline{c}_j & \text{if}\quad j\le n-p,\\ \\ w_j(\widetilde{u}(t)) & \text{if}\quad j>n-p,\end{cases}\qquad\forall\ t\ge 0\,, \tag{2.27}$$

for some constant values $\overline{c}_j\in[a_i,\,b_i]$, $j=1,\dots,n-p$, one has

$$E_t(\,\overline{u},\,\widetilde{u}\,)=E_t(\,\overline{u},\,\widetilde{u}'\,)\qquad\forall\ t\ge 0\,. \tag{2.28}$$

Therefore, given any $(\,\overline{u},\,\widetilde{u}\,),(\,\overline{v},\,\widetilde{v}\,)\in\mathcal{D}$, by replacing $\widetilde{u},\widetilde{v}$ in (2.25) with two auxiliary boundary data $\widetilde{u}',\widetilde{v}'$ having the property

$$w_j(\widetilde{u}'(t))=w_j(\widetilde{v}'(t))\qquad\forall\ t\ge 0,\qquad j=1,\dots,n-p,$$

we deduce for the flow map E_t the sharper estimate

$$\begin{aligned}&\big\|E_t(\,\overline{u},\,\widetilde{u}\,)\,-\,E_t(\,\overline{v},\,\widetilde{v}\,)\big\|_{\mathbf{L}^1([\delta,+\infty[)}\\ &\qquad\le L_t\cdot\left\{\big\|\,\overline{u}-\overline{v}\,\big\|_{\mathbf{L}^1(\mathbb{R}^+)}+\sum_{j=n-p+1}^{n}\big\|w_j(\widetilde{u})-w_j(\widetilde{v})\big\|_{\mathbf{L}^1([0,\,t])}\right\}.\end{aligned} \tag{2.29}$$

The next result states that every entropy weak solution to (1.1)–(1.3), admitting an essential limit in the $\mathbf{L}^1$ norm at time $t=0$, and at the boundary $x=0$, actually coincides with the corresponding trajectory $t\mapsto E_t(\,\overline{u},\,\widetilde{u}\,)$ of the flow map E_t constructed in Theorem 2.3. As a consequence, we deduce the uniqueness (up to the domain) of a Lipschitz continuous map as E_t having the property that each trajectory provide an entropy weak solution to (1.1)–(1.3) that admits a strong $\mathbf{L}^1$ trace at the boundary $x=0$, and at the initial time $t=0$. Notice that, in order to select a *unique* solution to (1.1)–(1.3), it is crucial to require that such a solution satisfies the decay estimates on positive waves stated in Definition 2.2(ii). For this reason, since in the case of $\mathbf{L}^\infty$ initial and boundary data the currently available results on the convergence of the viscous approximate solutions u^ε do not provide a priori BV bounds on u^ε, it remains an open problem whether or not the vanishing viscosity limit (2.21) coincides with the trajectory of the flow map E_t constructed in Theorem 2.3.

THEOREM 2.4. *Let* (1.1) *be a system of Temple class satisfying the same assumptions as in Theorem* 2.3. *Let* $u=u(t,x)$ *be an entropy weak solution to the mixed problem* (1.1)–(1.3) *on the region* $\Omega_T\doteq[0,\,T]\times\mathbb{R}^+$, *in the sense of Definition* 2.2. *Assume that the following conditions hold:*

(i) *The map* $(t,x)\to\big(u(t,\cdot),\,u(\cdot,x)\big)$ *takes values within the domain*

$$\mathcal{D}_T\doteq\Big\{\mathbf{p}=(\,\overline{u},\,\widetilde{u}\,)\ ;\ \overline{u}\in\mathbf{L}^1(\mathbb{R}^+,K),\,\widetilde{u}\in\mathbf{L}^1([0,T],K)\Big\}. \tag{2.30}$$

(ii) *For every fixed* $R>0$, *there holds*

$$\operatorname*{ess\,sup}_{t\to 0^+}\int_0^R\big|u(t,x)-\overline{u}(x)\big|\,dx=0\,. \tag{2.31}$$

(iii) *There holds*

$$\operatorname*{ess\,sup}_{x\to 0^+}\int_0^T\big|w_i\big(u(t,x)\big)-w_i(\widetilde{u}(t))\big|\,dt=0\qquad\forall\ i\in\{n-p+1,\dots,n\}\,. \tag{2.32}$$

Then u coincides with the corresponding trajectory of the flow map E_t, namely,

$$u(t,\cdot) = E_t(\,\overline{u},\,\widetilde{u}\,)(\cdot) \qquad \forall\ 0 \le t \le T. \tag{2.33}$$

A convenient way to prove that the regularity conditions (2.31)–(2.33) are verified is to employ the distributional entropy inequalities associated with the "boundary entropy pairs" for (1.1), as it is shown by Chen and Frid in [15, 16]. A pair of continuously differentiable functions $\alpha : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$, $\beta : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ is called a *boundary entropy pair* for (1.1) if, for any fixed $v \in \mathbb{R}^n$, $u \mapsto \big(\alpha(u,v),\,\beta(u,v)\big)$ is an entropy pair for (1.1) and there holds

$$\alpha(v,v) = \beta(v,v) = \partial_u \alpha(v,v) = 0 \qquad \forall\ v \in \mathbb{R}^n\,.$$

An immediate application of [15, Theorem 4.1] (or of [16, Theorem 1.1]) and of [15, Theorem 4.3] yields the following.

LEMMA 2.5. *Let $u(t,x)$ be an entropy weak solution to the mixed problem* (1.1)–(1.3) *on the region $\Omega_T \doteq [0,\,T] \times \mathbb{R}^+$, in the sense of Definition* 2.2. *Assume that, given any boundary entropy pair $\big(\alpha(u,v),\,\beta(u,v)\big)$ for* (1.1), *there is a constant $M > 0$ (depending only on $(\alpha,\,\beta)$ and on the domain K) such that, for every nonnegative test function $\phi \in \mathcal{C}^1_c(]-\infty, T[\times\mathbb{R}^+)$, and for any $v \in \mathbb{R}^n$, there holds*

$$\begin{aligned} &\int_0^T \int_0^{+\infty} \Big\{\alpha(u(t,x),\,v)\cdot\phi_t(t,x) + \beta(u(t,x),\,v)\cdot\phi_x(t,x)\Big\}\,dx\,dt \\ &\quad + \int_0^{+\infty} \big|\overline{u}(x) - v\big|\cdot\phi(0,x)\,dx + M\int_0^T \big|\widetilde{u}(t) - v\big|\cdot\phi(t,0)\,dt \ge 0\,. \end{aligned} \tag{2.34}$$

Then the essential limits (2.31)–(2.33) *are verified for any $R > 0$.*

Remark 2.6. By [15, Theorem 4.1] it also follows that, if $u(t,x)$ is an entropy weak solution to the mixed problem (1.1)–(1.3) on Ω_T in the sense of Definition 2.2, and if we assume that

(a) given any (standard) entropy pair $(\eta(u),\,q(u))$, for every test function $\phi \in \mathcal{C}^1_c(\overset{\circ}{\Omega}_T)$, $\phi \ge 0$, one has the usual entropy inequality

$$\int_0^T \int_0^{+\infty} \Big\{\eta(u(t,x))\cdot\phi_t(t,x) + q(u(t,x))\cdot\phi_x(t,x)\Big\}\,dx\,dt \ge 0, \tag{2.35}$$

(b) for every $R > 0$ there holds (2.31),

(c) given any boundary entropy pair $\big(\alpha(u,v),\,\beta(u,v)\big)$, for every function $\gamma \in \mathbf{L}^1([0,T])$, $\gamma \ge 0$ a.e., there holds

$$\operatorname*{ess\ sup}_{x\to 0^+} \int_0^T \beta(u(t,x),\,\widetilde{u}(t))\,\gamma(t)\,dt \le 0, \tag{2.36}$$

then the assumptions of Lemma 2.5 are verified, and hence the essential limits (2.33) hold.

2.3. Properties of the attainable sets. Following [4, 5], we now turn to study the mixed initial-boundary value problem (1.1)–(1.3) from the point of view of control theory, taking a fixed initial data $\overline{u} \in \mathbf{L}^1(\mathbb{R}^+, K)$ and considering, in connection with

a prescribed set $\mathcal{U} \subseteq \mathbf{L}^1(\mathbb{R}^+, K)$ of boundary data regarded as admissible controls, the *attainable sets* for (1.1)–(1.3):

(2.37)
$$\mathcal{A}(T, \mathcal{U}) \doteq \Big\{ E_T(\,\overline{u},\, \widetilde{u}\,)(\cdot)\ ;\ \widetilde{u} \in \mathcal{U} \Big\}, \qquad \mathcal{A}(\,\overline{x}, \mathcal{U}) \doteq \Big\{ E_{(\cdot)}(\,\overline{u},\, \widetilde{u}\,)(\overline{x})\ ;\ \widetilde{u} \in \mathcal{U} \Big\},$$

i.e., the sets of all profiles that can be attained at a fixed time $T > 0$, or at a fixed point in space $\overline{x} > 0$, by entropy weak solutions of (1.1)-(1.3) with initial data $\overline{u}$ and boundary data $\widetilde{u}$ that vary in $\mathcal{U}$. Relying on the well-posedness theory provided by the above results, we establish here the compactness of $\mathcal{A}(T, \mathcal{U})$, $\mathcal{A}(\,\overline{x}, \mathcal{U})$ for a class $\mathcal{U}$ of admissible boundary controls that satisfy convex constraints.

THEOREM 2.6. *Let K be a set of the form* (2.5) *and $J \subseteq \{1, \dots, n\}$ a set of indices such that $J \supseteq \{n-p, \dots, n\}$. Define*

(2.38)
$$\mathcal{U} \doteq \Big\{ \widetilde{u} \in \mathbf{L}^1(\mathbb{R}^+, K)\ ;\ w_j(\widetilde{u}(t)) \in [c_j,\ d_j], \qquad \textit{for a.e. } t \geq 0\,, \qquad \forall\ j \in J \Big\}$$

for some $-\infty < c_j \leq d_j < +\infty$, $j \in J$. Then $\mathcal{A}(T, \mathcal{U})$, $T > 0$, and $\mathcal{A}(\,\overline{x},\mathcal{U})$, $\overline{x} > 0$, are compact subsets of $\mathbf{L}^1{}_{loc}(\mathbb{R}^+, K)$.

3. Outline of the proof of Theorem 2.3. We describe here the basic steps in the proof of Theorem 2.3. All the technical estimates involved in the proof will then be worked out in sections 4–6. As in [14] we shall first construct a sequence of flow maps E^ν whose trajectories are front tracking approximate solutions [6, 10] of (1.1) in the region Ω that depend $\mathbf{L}^1$ continuously on the initial and boundary data. Next, for any fixed $\mathcal{M} > 0$, we shall prove the convergence of such a sequence of flow maps to a continuous flow of solutions $\mathbf{p} \mapsto E_t\,\mathbf{p}$, defined on the domain

$$\mathcal{D}_{\mathcal{M}} \doteq \Big\{ \mathbf{p} \in \mathcal{D};\ \text{Tot.Var.}\{\mathbf{p}\} \leq \mathcal{M} \Big\}, \tag{3.1}$$

where, if $\mathbf{p} = (\,\overline{u},\, \widetilde{u}\,)$,

$$\text{Tot.Var.}(\mathbf{p}) \doteq \text{Tot.Var.}(\overline{u}) + \text{Tot.Var.}(\widetilde{u}). \tag{3.2}$$

Finally, we will show that, for every fixed $\delta > 0$, and for any $t \geq \delta$, the map

$$\mathbf{p} \mapsto E_t\,\mathbf{p}\,\restriction_{[\delta, +\infty[}\,, \qquad \mathbf{p} \in \mathcal{D}_{\mathcal{M}} \tag{3.3}$$

is Lipschitz continuous with a Lipschitz constant depending on δ and t but independent of the bound on the total variation $\mathcal{M}$.

We now describe a front tracking algorithm which represents a natural extension of [14]. Fix an integer $\nu \geq 1$ and consider the discrete set of points in K whose coordinates are integer multiples of $2^{-\nu}$:

$$K^\nu \doteq \Big\{ u \in K\,;\ w_i(u) \in 2^{-\nu}\mathbb{Z}, \quad i = 1, \dots, n \Big\}.$$

Moreover, consider the domain

(3.4)
$$\mathcal{D}^\nu \doteq \Big\{ \mathbf{p} = (u, u') : \mathbb{R}^+ \mapsto K^\nu \times K^\nu; \qquad u,\, u' \in \mathbf{L}^1,\ \ u,\, u' \text{ are piecewise constant} \Big\}.$$

On $\mathcal{D}^\nu$ we now construct a flow map E^ν whose trajectories are front tracking approximate solutions of (1.1). To this end, we first describe how to solve a Riemann problem with left and right initial states $u^L, u^R \in K^\nu$. In Riemann coordinates, assume that

$$w(u^L) \doteq w^L = (w_1^L, \dots, w_n^L), \qquad w(u^R) \doteq w^R = (w_1^R, \dots, w_n^R).$$

Consider the intermediate states

$$z^0 = u^L, \quad \dots, \qquad z^i = u(w_1^R, \dots, w_i^R, w_{i+1}^L, \dots, w_n^L), \quad \dots, \quad z^n = u^R. \tag{3.5}$$

If $w_i^R < w_i^L$, the solution will contain a single i shock connecting the states z^{i-1}, z^i, and travelling with Rankine–Hugoniot speed $\lambda_i(z^{i-1}, z^i)$. Here and in what follows, by $\lambda_i(u, u')$ we denote the ith eigenvalue of the averaged matrix

$$A(u, u') \doteq \int_0^1 Df\big(\theta u + (1-\theta)u'\big)\, d\theta. \tag{3.6}$$

If $w_i^R > w_i^L$, the exact solution of the Riemann problem would contain a centered rarefaction wave. This is approximated by a rarefaction fan as follows. If $w_i^R = w_i^L + p_i\, 2^{-\nu}$ we insert the states

$$z^{i,\ell} = (w_1^R, \dots, w_i^L + \ell\, 2^{-\nu}, w_{i+1}^L, \dots, w_n^L), \qquad \ell = 0, \dots, p_i, \tag{3.7}$$

so that $z^{i,0} = z^{i-1}$, $z^{i,p_i} = z^i$. Our front tracking solution will then contain p_i fronts of the ith family, each connecting a couple of states $z^{i,\ell-1}$, $z^{i,\ell}$ and travelling with speed $\lambda_i\big(z^{i,\ell-1},\ z^{i,\ell}\big)$.

For a given pair of piecewise constant initial and boundary data $\mathbf{p} = (\,\overline{u},\, \widetilde{u}\,) \in \mathcal{D}^\nu$, the approximate solution $u(t,\cdot) \doteq E_t^\nu \mathbf{p}$ is now constructed as follows. At time $t = 0$, for $x > 0$ we solve each of the Riemann problems determined by the jumps in $\overline{u}$ according to the above procedure, while at $x = 0$ we construct the solution to the Riemann problem with left and right initial states $u^L = \widetilde{u}(0+)$, $u^R = \overline{u}(0+)$ and take its restriction to the interior of the domain Ω. This yields a piecewise constant function with finitely many fronts travelling with constant speeds. The solution is then prolonged up to the first time t_1 at which one of the following events takes place:

(a) Two or more discontinuities interact in the interior of Ω.
(b) One or more discontinuities hit the boundary.
(c) The boundary data $\widetilde{u}$ has a jump.

If case (a) occurs, then we solve the resulting Riemann problems, again applying the above procedure, while in cases (b)–(c) we construct the solution to the Riemann problem with left and right initial states $u^L = \widetilde{u}(t_1+)$, $u^R = u(t_1, 0+)$ and take its restriction to the interior of the domain Ω. This determines the solution $u(t,\cdot)$ until the time $t_2 > t_1$ where one of the events (a)–(c) again takes place, etc. Notice that at any time where case (b) occurs but (c) does not take place, no new wave is generated. Therefore, wave-fronts entering the domain Ω at the boundary $x = 0$ are produced only by the jumps of the boundary data $\widetilde{u}$.

As in [6] and [14], one checks that these front tracking approximations are well defined for all times $t \geq 0$. Indeed, the following properties hold:

- The total variation of $u(t,\cdot)$, measured w.r.t. the Riemann coordinates $w_1(t,\cdot), \dots, w_n(t,\cdot)$, is nonincreasing in time.
- The number of wave-fronts in $u(t,\cdot)$ is nonincreasing at each interaction. Hence, the total number of wave-fronts in $u(t,\cdot)$ remains finite.

It is now possible to define a ν-approximate semigroup $S^\nu : \mathbb{R}^+ \times \mathcal{D}^\nu \mapsto \mathcal{D}^\nu$ as in (2.24) by setting

$$S^\nu_t \mathbf{p} \doteq \big(E^\nu_t \mathbf{p},\, \mathcal{T}_t \widetilde{u}\big), \qquad \mathbf{p} = (\,\overline{u},\, \widetilde{u}\,). \tag{3.8}$$

The uniqueness of the definition of the approximate solution $u(t,\cdot) = E^\nu_t \mathbf{p}$ guarantees that S^ν_t satisfy the standard semigroup properties, i.e.,

$$S^\nu_0 = \text{Identity}, \qquad S^\nu_t \circ S^\nu_s = S^\nu_{t+s}.$$

Each trajectory $t \mapsto E^\nu_t \mathbf{p}$ is a weak solution of (1.1) (because all fronts satisfy the Rankine–Hugoniot conditions) but may contain discontinuities that do not satisfy the usual Lax stability conditions (because of the presence of rarefaction fronts).

We next proceed towards an estimate of the Lipschitz constant for $\mathbf{p} \mapsto E^\nu_t\, \mathbf{p}\restriction_{[\delta,+\infty[}$, $\delta > 0$, following the same technique adopted in [14]. The basic idea to estimate the distance between two approximate solutions u, v consists of constructing a continuous path of solutions u^θ connecting u, v and then studying how the length of the path $\theta \to u^\theta(t,\cdot)$ varies in time. In particular, given any two couples of initial and boundary data $\mathbf{p}_1 = (\,\overline{u}_1,\, \widetilde{u}_1\,)$, $\mathbf{p}_2 = (\,\overline{u}_2,\, \widetilde{u}_2\,)$ in $\mathcal{D}^\nu$, we introduce a suitable class of continuous paths (pseudopolygonals) that connect $\mathbf{fp}_1 \doteq (\,\overline{u}_1,\, f(\widetilde{u}_1))$ with $\mathbf{fp}_2 \doteq (\,\overline{u}_2,\, f(\widetilde{u}_2))$ by merely shifting the space and time positions of the jumps in $\overline{u}_1$, $\overline{u}_2$ and in $f(\widetilde{u}_1)$, $f(\widetilde{u}_2)$, respectively. More precisely, we consider a *pseudopolygonal* with values in

$$\mathcal{FD}^\nu \doteq \big\{\mathbf{fp} = (u, f(u')); \quad (u,u') \in \mathcal{D}^\nu\big\},$$

that is, a finite concatenation of *elementary paths* $\gamma : \theta \mapsto \big(\,\overline{u}^\theta,\, f(\widetilde{u}^\theta)\big)$ of the form

$$\begin{aligned} &\overline{u}^\theta(x) = \textstyle\sum_{\alpha=1}^N \overline{\omega}_\alpha \cdot \chi_{]x^\theta_{\alpha-1},\, x^\theta_\alpha]}(x), \qquad x^\theta_\alpha = x_\alpha + \xi_\alpha \theta, \quad x \geq 0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \theta \in [a,b], \\ &f(\widetilde{u}^\theta(t)) = \textstyle\sum_{\beta=1}^{\widetilde{N}} f(\widetilde{\omega}_\beta) \cdot \chi_{]t^\theta_{\beta-1},\, t^\theta_\beta]}(t), \qquad t^\theta_\beta = t_\beta + \widetilde{\xi}_\beta \theta, \quad t \geq 0, \end{aligned} \tag{3.9}$$

with $x^\theta_{\alpha-1} < x^\theta_\alpha$, $t^\theta_{\alpha-1} < t^\theta_\alpha$, for all $\theta \in [a,\, b]$ and $\alpha = 1, \ldots, N$, $\beta = 1, \ldots, \widetilde{N}$. Here, χ_I is the characteristic function of the interval I, while $\overline{\omega}_\alpha$, $\widetilde{\omega}_\beta \in K^\nu$ are constant states and ξ_α, $\widetilde{\xi}_\beta$ are, respectively, the (space) shift rate of the jump in $\overline{u}^\theta$ at x_α and the (time) shift rate of the jump in $f(\widetilde{u}^\theta)$ at t_β. A simple example of a pseudopolygonal joining two couples of initial data and boundary flux $\mathbf{fp}_1 = (\,\overline{u}_1,\ f(\widetilde{u}_1))$, $\mathbf{fp}_2 = (\,\overline{u}_2,\ f(\widetilde{u}_2))$ is given by

$$\theta \mapsto \big(\,\overline{u}_1 \cdot \chi_{[0,\theta[} + \overline{u}_2 \cdot \chi_{]\theta,+\infty[}\,,\ \ f(\widetilde{u}_1) \cdot \chi_{[0,\theta[} + f(\widetilde{u}_2) \cdot \chi_{]\theta,+\infty[}\,\big).$$

The $\mathbf{L}^1$ length of an elementary path γ of the form (3.9) is then computed by

$$\begin{aligned} \|\gamma\|_{\mathbf{L}^1} &= \int_a^b \left\{ \sum_{\alpha=1}^N \big|\Delta\overline{u}^\theta(x_\alpha)\big| \left|\frac{\partial x^\theta_\alpha}{\partial\theta}\right| + \sum_{\beta=1}^{\widetilde{N}} \big|\widetilde{\Delta}\widetilde{u}^\theta(t_\beta)\big| \left|\frac{\partial t^\theta_\beta}{\partial\theta}\right| \right\} d\theta \\ &= \left\{ \sum_{\alpha=1}^N \big|\sigma_\alpha\big|\,|\xi_\alpha| + \sum_{\beta=1}^{\widetilde{N}} \big|\widetilde{\sigma}_\beta\big|\,|\widetilde{\xi}_\beta| \right\} (b-a), \end{aligned} \tag{3.10}$$

where

$$\sigma_\alpha \doteq \Delta\overline{u}^\theta(x_\alpha) = \overline{\omega}_{\alpha+1} - \overline{\omega}_\alpha, \qquad \widetilde{\sigma}_\beta \doteq \widetilde{\Delta}\widetilde{u}^\theta(t_\beta) = f(\widetilde{\omega}_{\beta+1}) - f(\widetilde{\omega}_\beta). \tag{3.11}$$

If we consider a pseudopolygonal $\gamma_0^\nu : \theta \mapsto \big(\overline{u}^\theta,\, f(\widetilde{u}^\theta)\big)$, $\theta \in [0,\,1]$, with values in $\mathcal{FD}^\nu$, and let $u_\nu^\theta(t,\cdot) = E_t^\nu(\,\overline{u}^\theta,\,\widetilde{u}^\theta)$ be the corresponding solution, since the number of wave-fronts in these solutions is a priori bounded and the locations of the interaction points in the t-x plane are determined by a linear system of equations, it follows that, at any time $t > 0$, the path

$$\gamma_t^\nu : \theta \mapsto \big(u_\nu^\theta(t,\cdot),\, f(\widetilde{u}^\theta)\big), \qquad \theta \in [0,\,1], \tag{3.12}$$

is still a pseudopolygonal with values in $\mathcal{FD}^\nu$. Moreover, there exist finitely many parameter values $0 = \theta_0 < \theta_1 < \cdots < \theta_m = 1$ such that the wave-front configuration of u_ν^θ remains the same as θ ranges on each of the open intervals $I_j \doteq]\theta_{j-1}, \theta_j[$. In this case, the length of the path γ_t^ν is measured by an expression of the form

$$\begin{aligned}\|\gamma_t^\nu\|_{\mathbf{L}^1} =& \sum_{j=1}^m \int_{\theta_{j-1}}^{\theta_j} \sum_\alpha \big|\Delta u_\nu^\theta(t,\, x_\alpha^\theta)\big| \left|\frac{\partial x_\alpha^\theta}{\partial\theta}\right| d\theta \\ &+ \sum_{j=1}^m \int_{\theta_{j-1}}^{\theta_j} \sum_\beta \big|\widetilde{\Delta}\widetilde{u}^\theta(t_\beta^\theta)\big| \left|\frac{\partial t_\beta^\theta}{\partial\theta}\right| d\theta\,.\end{aligned} \tag{3.13}$$

Let $\pi_1(\,\overline{u},\, f(\widetilde{u})) = \overline{u}$, $\pi_2(\,\overline{u},\, f(\widetilde{u})) = f(\widetilde{u})$, and denote the canonical projections for any couple $\mathbf{fp} = (\,\overline{u},\, f(\widetilde{u})) \in \mathcal{FD}^\nu$. In connection with any elementary path γ of the form (3.9), define the paths

$$\rho^i_{s_1,s_2}(\gamma) \;:\; \theta \longmapsto \pi_i(\gamma(\theta)) \restriction_{[s_1,\,s_2]}, \qquad s_1,\, s_2 \geq 0\,, \tag{3.14}$$

and introduce the seminorms

$$\big\|\gamma\big\|_{\delta,t_1,t_2} \doteq \big\|\rho^1_{\delta,+\infty}(\gamma)\big\|_{\mathbf{L}^1} + \big\|\rho^2_{t_1,t_2}(\gamma)\big\|_{\mathbf{L}^1}, \qquad \delta,\, t_i \geq 0. \tag{3.15}$$

Since the second term of the sum in (3.13) is constant in time, we have

$$\big\|\gamma_t^\nu\big\|_{\delta,0,t} = \big\|\rho^1_{\delta,+\infty}(\gamma_t^\nu)\big\|_{\mathbf{L}^1} + \big\|\rho^2_{0,t}(\gamma_0^\nu)\big\|_{\mathbf{L}^1} \qquad \forall\; t \geq 0\,. \tag{3.16}$$

Thus, to estimate the $\mathbf{L}^1$ distance between two approximate solutions $E_t^\nu\mathbf{p}_1$, $E_t^\nu\mathbf{p}_2$, we will provide in Lemma 5.1 an a priori bound on the integrand of the first term in (3.13), which represents the infinitesimal length of a generalized tangent vector to the one-parameter family of pairs of solutions and boundary flux (3.12). Relying on this result, we will show in section 6.1 that, for any fixed $\mathcal{M} > 0$, $\delta > 0$, and for any $t \geq \delta$, there exists some constant $L_{\mathcal{M},t} \doteq c_0(1+\mathcal{M})(1+\log(t/\delta)) > 0$ such that there holds

$$\big\|\rho^1_{\delta,+\infty}(\gamma_t^\nu)\big\|_{\mathbf{L}^1} \leq L_{\mathcal{M},t}\cdot\big\|\gamma_0^\nu\big\|_{0,0,t} \qquad \forall\; t \geq \delta \tag{3.17}$$

for every pseudopolygonal $\gamma_0^\nu : [0,\,1] \to \mathcal{FD}^\nu$ joining two couples of initial data and boundary flux in $\{\mathbf{fp};\;\; \mathbf{p} \in D_\mathcal{M} \cap \mathcal{D}^\nu\}$. Hence, defining the pseudometrics

$$\begin{aligned} d_{\delta,t_1,t_2}(\mathbf{p}_1,\mathbf{p}_2) \doteq& \big\|\,\overline{u}_1 - \overline{u}_2\big\|_{\mathbf{L}^1([\delta,\,+\infty[)} + \big\|\,\widetilde{u}_1 - \widetilde{u}_2\,\big\|_{\mathbf{L}^1([t_1,\,t_2])}, \\ &\mathbf{p}_i = (\,\overline{u}_i,\,\widetilde{u}_i\,), \quad \delta,\, t_i \geq 0, \end{aligned} \tag{3.18}$$

and observing that the $\mathbf{L}^1$ lengths of the paths γ_0^ν, γ_t^ν satisfy

$$\|\gamma_0^\nu\|_{0,0,t} \leq C_0\cdot d_{0,0,t}(\mathbf{fp}_1,\,\mathbf{fp}_2), \tag{3.19}$$

$$\big\|E_t^\nu\mathbf{p}_1 - E_t^\nu\mathbf{p}_2\big\|_{\mathbf{L}^1([\delta,+\infty[)} \leq C_0\cdot\|\gamma_t^\nu\|_{\delta,0,t}\,, \tag{3.20}$$

for some constant $C_0 > 0$ (depending only on the domain $\mathcal{D}$), we deduce from (3.16)–(3.17) a uniform Lipschitz estimate for the flow maps $\mathbf{p} \mapsto E^\nu_t\,\mathbf{p}\restriction_{[\delta,+\infty[}$ of the type

$$\big\|E^\nu_t\mathbf{p}_1 - E^\nu_t\mathbf{p}_2\big\|_{\mathbf{L}^1([\delta,+\infty[)} \leq L'_{\mathcal{M},t}\cdot d_{0,0,t}(\mathbf{fp}_1,\,\mathbf{fp}_2) \qquad \forall\ t\geq\delta\,, \tag{3.21}$$

with $L'_{\mathcal{M},t} \doteq c'_0(1+\mathcal{M})(1+\log(t/\delta))$, for some other constant $c'_0 > 0$ independent of ν, and for any $\mathbf{p}_1, \mathbf{p}_2 \in D_{\mathcal{M}} \cap \mathcal{D}^\nu$. As $\nu\to\infty$, the domain $D_{\mathcal{M}}\cap\mathcal{D}^\nu$ become dense in $\mathcal{D}_{\mathcal{M}}$. In the limit, a continuous flow map E_t is obtained in (6.2), defined on the domain $\mathcal{D}_M$ and satisfying the estimate (2.25).

To extend the flow map E_t to the whole domain $\mathcal{D}$ preserving the property (2.25), by similar arguments as above we will prove in section 6.2 the estimate

$$\big\|E_t\mathbf{p}_1 - E_t\mathbf{p}_2\big\|_{\mathbf{L}^1([\delta,+\infty[)} \leq L''_t\cdot d_{0,0,t}(\mathbf{fp}_1,\,\mathbf{fp}_2) \qquad \forall\ t\geq\delta \tag{3.22}$$

with $L''_t \doteq c''_0(1+\log(t/\delta))$ for some other constant $c''_0 > 0$ independent of the total variation, and for any $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{D}^\mu$, $\mu \geq 1$. Any trajectory $t\mapsto E_t(\mathbf{p})$ of such a map E_t (defined in (6.12)) is defined as the limit of front tracking approximations, and hence provides a weak solution to problem (1.1)–(1.2).

In order to show that the map $(t,x)\mapsto E_t(\,\overline{u},\,\widetilde{u}\,)(x)$ admits a strong $\mathbf{L}^1$ trace at the boundary $x=0$, we next derive a stability estimate for the map $\mathbf{p}\mapsto f\big(E_{(\cdot)}\mathbf{p}(x)\big)$ following the same homotopy and linearization technique adopted above. Namely, we first establish in Lemma 5.2 an a priori bound on a generalized tangent vector to the one-parameter family of pairs of initial data and fluxes

$$\gamma^\nu_x : \theta\mapsto \big(\,\overline{u}^\theta,\, f\big(u^\theta_\nu(\cdot,\,x)\big)\big), \qquad \theta\in[0,\,1], \tag{3.23}$$

evaluated along the vertical segment of the domain Ω. Next, we show in section 6.3 that, for any $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{D}^\mu,\ \mu\geq 1$, and for every $\tau_2 > \tau_1 > 0$, there holds

$$\begin{aligned}&\big\|f\big(E_{(\cdot)}\mathbf{p}_1(x)\big) - f\big(E_{(\cdot)}\mathbf{p}_2(x)\big)\big\|_{\mathbf{L}^1([\tau_1,\,\tau_2])} \leq L'''\cdot d_{0,0,\tau_2}(\mathbf{fp}_1,\,\mathbf{fp}_2)\\ &\qquad\qquad \forall\ x\in[0,\,(\lambda^{\min}/2)\,\tau_1]\,,\end{aligned} \tag{3.24}$$

where $\lambda^{\min}$ is the lower bound for the absolute value of all characteristic speeds in (2.7), and $L''' \doteq c'''_0(1+\log(\tau_2/\tau_1))$, for some constant $c'''_0 > 0$ independent of μ. By continuity, and relying on the density of the domains $\mathcal{D}^\mu,\ \mu\geq 1$ in $\mathcal{D}$, we then extend the estimate (3.24) to any pair $\mathbf{p}_1,\ \mathbf{p}_2$ in $\mathcal{D}$. Relying on this property, and thanks to the invertibility of the flux f (see Remark 2.1), we prove in section 6.5 the existence of the strong $\mathbf{L}^1$ trace of $E_t\mathbf{p}(x)$ at $x=0$ for any $\mathbf{p}\in\mathcal{D}$. Finally, we show in (6.24) that $E_t\mathbf{p}$ fulfills the boundary condition (2.20), and we prove in section 6.4 that the Oleinik-type estimates (2.15)–(2.18) on the decay of the positive waves are satisfied, thus completing the proof of Theorem 2.3.

4. Preliminary results. Fix $\nu\geq 1$ and consider a piecewise constant solution $u(t,\cdot\,)\doteq E^\nu_t(\,\overline{u},\,\widetilde{u}\,)$ of (1.1) in the region Ω constructed by the front tracking algorithm described in section 3 for some $(\,\overline{u},\,\widetilde{u}\,)\in\mathcal{D}^\nu$. We may perturb this solution by shifting the (space) locations x_α of the jumps in the initial data $\overline{u}$ at rates ξ_α and the (time) locations t_α of the jumps in the boundary data $\widetilde{u}$ entering the interior of the domain Ω at rates $\widetilde{\xi}_\alpha$ (Figure 1). This means that, if we let $x_\beta = x_\beta(t)$ denote the jumps in the unperturbed solution $u(t,\cdot)$, for θ suitably close to zero, the corresponding perturbed solution $u^\theta(t,\cdot)$ will be a function with jumps at the points $x^\theta_\beta = x_\beta + \theta\xi_\beta$. In the same way, $u^\theta(\cdot,x)$ will have jumps at times $t^\theta_\beta = t_\beta - \theta\widetilde{\xi}_\beta$ with $\widetilde{\xi}_\beta = \xi_\beta/\lambda_{k_\beta}$, where $t_\beta = t_\beta(x)$

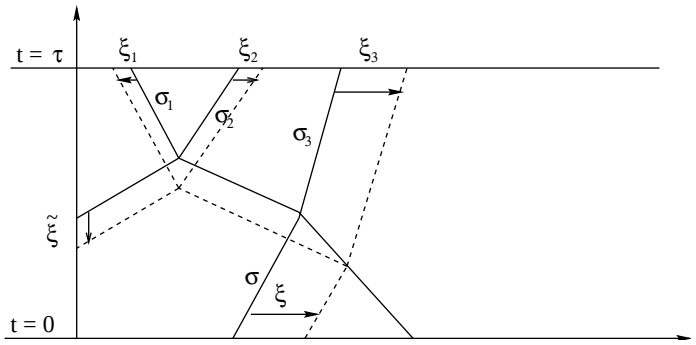


FIG. 1.

denote the locations of all jumps in $u(\cdot, x)$ having nonzero slope λ_{k_β}. As long as the wave-front configuration of the functions u, u^θ is the same, the *space-shifts* $\xi_\beta(t)$ and the *time-shifts* $\widetilde{\xi}_\beta(x)$ are uniquely determined as linear functions of the initial time and space-shifts ξ_α, $\widetilde{\xi}_\alpha$. In this section we collect some basic properties of these *shift differentials* that depend on the special geometric features of Temple class systems and on the fact that the front tracking algorithm described in section 3 guarantees that wave-fronts entering the domain Ω at the boundary $x = 0$ are produced only by the jumps of the boundary data. Such properties can be obtained with entirely similar arguments as for the corresponding results in [14]. Hence, we refer to [14] for most of the proofs of the results presented in this section, limiting ourself to discussing the points that really involve the boundary conditions and to establish in detail the proof of Lemma 4.4 which provides decay estimates on the positive waves of front tracking solutions that are different from the corresponding ones (Lemmas 4 and 5) in [14].

Remark 4.1. We denote by $\sigma_\alpha(t) = u(t, x_\alpha+) - u(t, x_\alpha-)$ the size of a jump in the solution u, occurring at $(t, x_\alpha(t))$, along the space direction (*space-size*) and by $\widetilde{\sigma}_\alpha(x) = f(u(t_\alpha+, x)) - f(u(t_\alpha-, x))$ the size of a jump in the flux $f(u)$ occurring at $(t, x_\alpha(t))$ along the time direction (*time-size*). Since approximate solutions are indeed weak solutions, by the Rankine–Hugoniot equations we have the identity

$$\widetilde{\xi}_\alpha \widetilde{\sigma}_\alpha = \frac{\xi_\alpha}{\lambda_{k_\alpha}} \lambda_{k_\alpha} \sigma_\alpha = \xi_\alpha \sigma_\alpha. \tag{4.1}$$

In the following we will use both notations, depending on convenience.

LEMMA 4.1. *Consider a bounded, open region Γ in Ω. Call σ_α, $\alpha = 1, \dots, M$, the fronts entering Γ and let ξ_α be their space-shifts. Assume that the fronts leaving Γ, say σ_β, $\beta = 1, \dots, M'$, are linearly independent. Then the space-shifts ξ_β of the outgoing fronts are uniquely determined by the linear relation*

$$\sum_{\alpha=1}^{M} \xi_\alpha \sigma_\alpha = \sum_{\beta=1}^{M'} \xi_\beta \sigma_\beta \,. \tag{4.2}$$

A proof of Lemma 4.1, based on an application of the divergence theorem, can be obtained by the same arguments in [14].

Remark 4.2. As observed in [14], according to Lemma 4.1 the shift rates of the outgoing fronts from a given region Γ depend only on the shift rates of the incoming ones and not on the order in which these wave-fronts interact inside Γ. In particular,

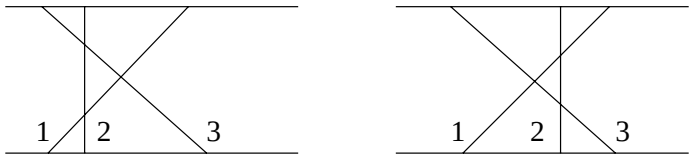


FIG. 2.

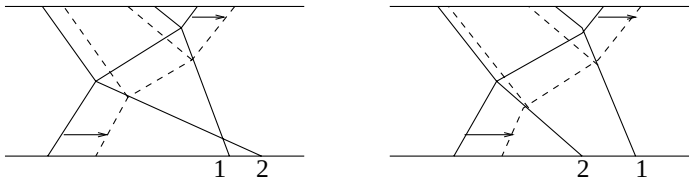


FIG. 3.

one can perform the following two operations, without changing the shift rates of the outgoing fronts:

(O1) Switch the order in which three fronts interact (Figure 2).

(O2) Invert the order of two fronts at $t = 0$ or at $x = 0$, provided that both fronts have zero shift rate (Figure 3).

This property will be used repeatedly in our future estimates. Indeed, in the computation of a shift rate, we can suitably alter the order of wave interactions and thus reduce the problem to a case where the wave-front configuration is particularly simple.

LEMMA 4.2. *Assume that a front tracking solution* $u(t,\cdot) = E^\nu_t(\overline{u}, \widetilde{u})$ *contains two wave-fronts of the same characteristic family, say* $t \mapsto x_{\alpha'}(t)$, $t \mapsto x_{\alpha''}(t)$, *originating at distinct points* $(\tau', \overline{x}')$, $(\tau'', \overline{x}'')$, $\tau' \geq \tau''$, $\overline{x}' \leq \overline{x}''$, *of the boundary of* Ω, *and such that* $x_{\alpha'}(t) \leq x_{\alpha''}(t)$, $t \in [\tau', T]$. *Then it is possible to assign space-shift rates* ξ_α *to all fronts in the initial data* $\overline{u}$ *and in the boundary data* $\widetilde{u}$ *so that the space-shift rate of the front at* $x_{\alpha'}(\tau', \overline{x}')$ *is* $\xi_{\alpha'} = 1$, *and, in the corresponding perturbed solution* u^θ, *all fronts* $x_\beta(t)$ *outside the strip* $\Gamma \doteq \big\{(t,x);\ \ t \in [\tau'', \tau'],\ 0 \leq x \leq x_{\alpha''}(t)\big\} \cup \big\{(t,x);\ \ t \in [\tau',\ T],\ x_{\alpha'}(t) \leq x \leq x_{\alpha''}(t)\big\}$ *have zero shift rate.*

In other words, the perturbation of the initial and boundary data can be chosen so that one particular front shifts at unit rate, but the corresponding solution remains unaffected outside the region Γ (Figure 4). For a proof of the lemma, proceed as in [14].

LEMMA 4.3. *Let* u *be a front tracking solution of* (1.1) *in the region* Ω *and consider two wave-fronts, say* $t \mapsto x(t)$, $t \in [\tau_x, T]$, *and* $t \mapsto y(t)$, $t \in [\tau_y, T]$, *originating at two points* $(\tau_x, \overline{x})$, $(\tau_y, \overline{y})$ *of the boundary of* Ω. *Then there exists a second front tracking solution* $\widehat{u}$ *with two fronts* $t \mapsto \widehat{x}(t)$, $t \in [\tau_x, T]$, $t \mapsto \widehat{y}(t)$, $t \in [\tau_y, T]$, *having the following properties:*

(i) $\widehat{x}(\tau_x) = \overline{x}$, $\widehat{y}(\tau_y) = \overline{y}$, $\widehat{x}(T) = x(T)$, $\widehat{y}(T) = y(T)$.

(ii) $\widehat{u} = u$ *in a neighborhood of these points* $(\tau_x, \overline{x})$, $(\tau_y, \overline{y})$, $(T, x(T))$, $(T, y(T))$.

(iii) *Tot.Var.*$\big\{\big(\widehat{u}(0,\cdot), \widehat{u}(\cdot, 0)\big)\big\} \leq C_1$ *for some constant* C_1 *depending only on the system* (1.1) *and on the set* K.

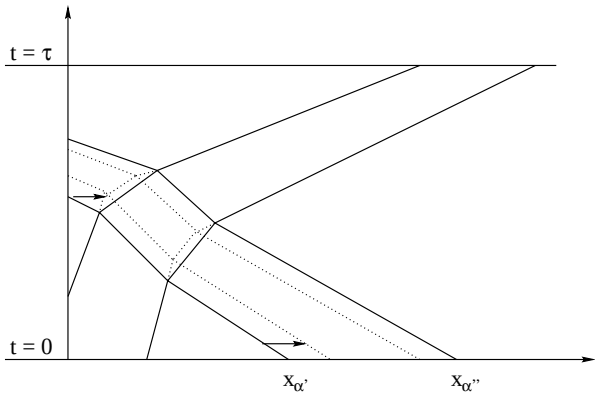


FIG. 4.

Proof. To fix the ideas, assume $\overline{x} = 0$, $\tau_x > 0$, $\overline{y} > 0$, $\tau_y = 0$. Call J_1, J_2, J_3 the three connected components of the set $\big(\{0\}\times\mathbb{R}^+\cup[0,\ T]\times\{0\}\big)\setminus\{(0,\overline{y}\,),\ (\tau_x,0)\}$ and, similarly, let J_1', J_2', J_3', J_4' denote the connected components of the set $\big([0,T]\times\{0\}\cup\{T\}\times\mathbb{R}^+\big)\setminus\{(\tau_x,0),\ (T,x(T)),\ (T,y(T))\}$. Assume that u contains two wave-fronts $t\mapsto z'(t)$, $t\mapsto z''(t)$, of the same kth family and with the same sign, starting at some points ζ_I', ζ_I'' within the same set J_i, and ending at some other points ζ_F', ζ_F'' that lie in the same set J_j'. Then, proceeding as in [14], we may apply Lemma 4.2 and obtain a second front tracking solution u^{θ_1} which has a smaller number of k-fronts than u and coincides with u outside the region bounded by the fronts z', z''. We can repeat this construction as long as the resulting solution contains fronts of the same family and with the same sign, starting at points of the same set J_i and ending within the same set J_j'. In a finite number of steps, we then obtain a new solution $\widehat{u}$ which has the property that, for each $k = 1,\dots,n$, and for any $i = 1,2,3,\quad j = 1,2,3,4$, there exists at most one point $\zeta^0\in J_i$ where a positive k-wave originates, terminating within J_j', and similarly for negative k-waves. This implies that the total variation of $(\widehat{u}(0,\cdot),\,\widehat{u}(\cdot,0)\restriction_{[0,T]})$ is uniformly bounded by a constant C_1 depending only on n and on the diameter of the set K, which completes the proof of the lemma, since we may clearly modify $\widehat{u}(t,\cdot)$ for $t > T$ so that also $\text{Tot.Var.}\big\{\widehat{u}(\cdot,0)\restriction_{[T,+\infty]}\big\}\le C_1$. □

Due to genuine nonlinearity, the amount of positive waves in $u(t,\cdot)$ contained in an interval $[a,\ b]$, measured in Riemann coordinates, decays in time. We have the following result.

LEMMA 4.4. *Consider a front tracking solution $u(t,\cdot) = E_t^{\nu}(\,\overline{u},\,\widetilde{u}\,)$, with $\overline{u}$, $\widetilde{u}$ containing together at most N shock fronts of the kth family. Then there exists some constant C_2 depending only on the system* (1.1) *such that, for each $\tau > 0$, and for every interval $[a,\ b]$, $a > 0$, one has*

$$\text{Tot.Var.}\big\{w_k(\tau,\cdot);\ [a,\ b]\big\}\le 2C_2\frac{b-a}{\tau}+\|w_k\|_{L^\infty}+(N+1)2^{1-\nu} \tag{4.3}$$

for $k = 1,\dots,n-p$,

$$\text{Tot.Var.}\big\{w_k(\tau,\cdot);\ [a,\ b]\big\}\le 2C_2\left\{\frac{b-a}{\tau}+\log\frac{b}{a}\right\}+\|w_k\|_{L^\infty}+(N+1)2^{1-\nu} \tag{4.4}$$

for $k = n-p+1,\dots,n$.

Proof. We give the proof of the statement only for $k \in \{n-p+1, \dots, n\}$, the other case being entirely similar. Relying on Lemma 4.3 and on the uniform strict hyperbolicity assumption (SH1), with the same arguments used in the proof of Lemma 4 in [14] one can show that any two adjacent k-rarefaction fronts $x(t) \leq y(t)$ of u, starting from the boundary $x = 0$, are separated at time $\tau > 0$ by a distance $\geq \kappa(\tau - t_0) \cdot 2^{-\nu}$, where $t_0 \geq 0$ is the beginning time of the rarefaction front $x(t)$, and $\kappa > 0$ denotes some constant depending only on the system. Hence, the distance between rarefaction fronts entering the domain Ω from the boundary $x = 0$ grows at least linearly with the distance from the t-axis. Therefore, the number of rarefaction fronts emanating from the boundary and crossing any interval $[a, b]$, $a > 0$, is bounded by

$$1 + N + \frac{C_2}{2^{-\nu}} \log \frac{b}{a}$$

for some constant $C_2 > 0$ depending on the system (1.1). The positive variation of $w_k(\tau, \cdot)$ on $[a, b]$, i.e., the total amount of upward jumps, thus satisfies

$$\text{Pos.Var.}\big\{w_k(\tau, \cdot);\ [a,\ b]\big\} \leq (1+N)2^{-\nu} + C_2 \cdot \log \frac{b}{a}. \tag{4.5}$$

On the other hand, the same decay estimates in [14, Lemma 5] hold for the waves starting from $t = 0$:

$$\text{Pos.Var.}\big\{w_k(\tau, \cdot);\ [a,\ b]\big\} \leq (1+N)2^{-\nu} + C_2 \cdot \frac{b-a}{\tau}\,. \tag{4.6}$$

In turn, the total variation of $w_k(\tau, \cdot)$ on $[a, b]$ is bounded by $\|w_k\|_{\mathbf{L}^\infty}$ plus twice the positive variation of w_k. Hence (4.3)–(4.4) hold. □

5. Estimates on shift differentials. In this section, relying on the results presented in section 4, we recover the key estimates on shift differentials of paths of approximate solutions. We will use the same technique developed in [14], since we are dealing with front tracking solutions whose wave-fronts emanating from the boundary are produced only by the jumps on the boundary data.

LEMMA 5.1. *Let $u(t, \cdot) = E_t^\nu(\,\overline{u},\, \widetilde{u}\,)$ be a front tracking solution, with $\overline{u}$, $\widetilde{u}$ containing together N shocks. Assume that the fronts of $\overline{u}$ located at x_β (respectively, the fronts of $f(\widetilde{u})$ starting at t_β) are shifted with shift rate ξ_β (respectively, $\widetilde{\xi}_\beta = \xi_\beta/\lambda_{k_\beta}$) and have amplitude σ_β (respectively, $\widetilde{\sigma}_\beta = \lambda_{k_\beta}\,\sigma_\beta$). Then there exists a constant C_3 depending only on the system* (1.1) *and on the domain K such that, for any $\delta > 0$, and for every $\tau \geq \delta$, calling $\xi_\alpha(\tau)$, $\sigma_\alpha(\tau)$ the shift rates and the amplitudes of the fronts in $u(\tau, \cdot)$, we have*

$$\sum_{x_\alpha(\tau) > \delta} \big|\xi_\alpha(\tau)\,\sigma_\alpha(\tau)\big| \leq C_3(1 + N2^{-\nu})(1 + \log(\tau/\delta)) \cdot \left(\sum_\beta \big|\xi_\beta\,\sigma_\beta\big| + \sum_\beta \big|\widetilde{\xi}_\beta\,\widetilde{\sigma}_\beta\big|\right). \tag{5.1}$$

Proof. Assume first that only one single front σ^0 is shifted, starting at time $t = 0$ in the position $x = \overline{x}$ (or leaving the boundary $x = 0$ at time $t = \overline{t}$), say, of the kth family, with shift rate ξ^0. Consider one particular front, say, located at $x_{\alpha^*}(\cdot)$, of the jth family, and call $\overline{y} \doteq x_{\alpha^*}(\tau)$ its terminal point at time τ. We claim that there are

constants C_4, C_5, depending only on the system (1.1) and on the domain K, such that the following properties hold.

(P1) If x_{α^*} is precisely the k-front starting at $\overline{x}$ ($\overline{t}$, respectively), then

$$\left|\xi_{\alpha^*}(\tau)\,\sigma_{\alpha^*}(\tau)\right| \leq C_4|\xi^0\,\sigma^0|. \tag{5.2}$$

(P2) If x_{α^*} is a j-front, with $j \neq k$, and the backward j-characteristics ending at $\overline{y}$ include fronts starting from both sides of $\overline{x}$ ($\overline{t}$, respectively), then (5.2) again holds.

(P3) If x_{α^*} is a j-front, and the j-fronts ending at $\overline{y}$ start all at the same side of $\overline{x}$ ($\overline{t}$, respectively), one then has the sharper estimate

$$\left|\xi_{\alpha^*}(\tau)\right| \leq C_5|\xi^0\,\sigma^0|. \tag{5.3}$$

Properties (P1)–(P2) can be established by the same arguments in [14], with minor changes. Hence, we limit ourselves here to give a proof of (P3). To this end, observe that, besides the fronts starting at $\overline{x}$ (or $\overline{t}$) and the ones ending at $\overline{y}$, one can single out four groups of waves:

(1) the waves starting on the left of $\overline{x}$ (respectively, after $\overline{t}$) and ending on the left of $\overline{y}$;
(2) the waves starting on the right of $\overline{x}$ (respectively, before $\overline{t}$) and ending on the right of $\overline{y}$;
(3) the waves starting on the right of $\overline{x}$ (respectively, before $\overline{t}$) and ending on the left of $\overline{y}$;
(4) the waves starting on the left of $\overline{x}$ (respectively, after $\overline{t}$) and ending on the right of $\overline{y}$.

According to Remark 4.2, in our computation of the shift rate $\xi_{\alpha^*}(\tau)$ of the front reaching $\overline{y}$, it is not restrictive to assume that the sets of waves in (1) and (2) are empty. Indeed, we can otherwise shift the locations of all these fronts of type (1) towards the left, until they all lie outside the domain influenced by the shift at $\overline{x}$. Similarly, fronts of type (2) can be shifted toward the right until they lie completely outside this domain of influence.

Having achieved this simplification, we shall first establish (P3) in the case (Figure 5) where no j-wave ending at $\overline{y}$ crosses the k-wave starting at $\overline{x}$ (or $\overline{t}$). Consider a curve γ running slightly to the right of the minimal backward j-front ending at $\overline{y}$. By

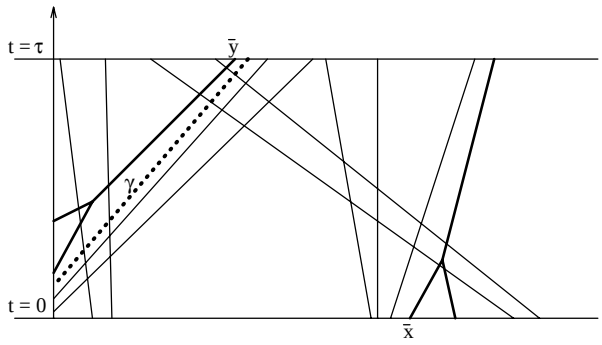


FIG. 5.

Lemmas 4.1 and 4.2, after performing the operations (O1)–(O2) a number of times, we can consider an equivalent configuration with the following properties:

- No front crosses γ from left to right.
- There exists some index $\ell \leq j$ such that only fronts of families $i < \ell$ can cross γ from right to left, and we can assume that the waves of type (1) have zero shift rate at every time in the interval $[0,\ \tau]$.

Applying Remark 4.2 to the region on the right of γ we obtain

$$\sum_{\alpha\in C(\gamma)} \xi_\alpha\,\sigma_\alpha + \sum_{x_\alpha(\tau)>\gamma(\tau)} \xi_\alpha(\tau)\,\sigma_\alpha(\tau) = \xi^0\,\sigma^0. \tag{5.4}$$

Here the first summation extends to all fronts crossing the curve γ with nonzero shift rate. Call u^L and u^R the left and right states across the jump at $\overline{y}$. Observing that the two sums on the left-hand side of (5.4) are contained in

$$\text{span}\big\{r_1(u^R),\dots,r_{\ell-1}(u^R)\big\},\qquad \text{span}\big\{r_\ell(u^R),\dots,r_n(u^R)\big\}, \tag{5.5}$$

using the strict hyperbolicity condition (SH2), we conclude that

$$\left|\sum_{\alpha\in C(\gamma)} \xi_\alpha\sigma_\alpha\right| \leq C'\left|\xi^0\,\sigma^0\right| \tag{5.6}$$

for some constant C', depending only on the system (1.1). We now again apply Remark 4.2 to the region on the left of γ. Observing that the only incoming fronts which carry a nonzero shift rate are those crossing γ from right to left, and that the only outgoing shifted j-front is the one ending at $\overline{y}$, we obtain

$$\sum_{x_\alpha(\tau)<\overline{y}} \xi_\alpha\sigma_\alpha + \xi_{\alpha^*}(\tau)\sigma_{\alpha^*}(\tau) = \sum_{\alpha\in C(\gamma)} \xi_\alpha\sigma_\alpha. \tag{5.7}$$

Recalling the normalization at (2.1), we observe that (5.7) implies

$$\big|\xi_{\alpha^*}(\tau)\sigma_{\alpha^*}(\tau)\big| = l_j(u^L)\cdot\sum_{\alpha\in C(\gamma)} \xi_\alpha\sigma_\alpha. \tag{5.8}$$

On the other hand, one has

$$l_j(u^R)\cdot\sum_{\alpha\in C(\gamma)} \xi_\alpha\sigma_\alpha = 0. \tag{5.9}$$

Together, (5.6), (5.8) and (5.9) imply

$$\big|\xi_{\alpha^*}(\tau)\sigma_{\alpha^*}(\tau)\big| \leq \big|l_j(u^R)-l_j(u^L)\big|\left|\sum_{\alpha\in C(\gamma)} \xi_\alpha\sigma_\alpha\right| \leq C_5\big|\sigma_{\alpha^*}(\tau)\big|\big|\xi^0\,\sigma^0\big| \tag{5.10}$$

for some other constant C_5 depending only on the system (1.1), proving (5.3).

We next establish (P3) in the case where $k > j$ and all j-waves running into $\overline{y}$ cross the k-wave starting from $\bar{t}$, as in Figure 6. In this case, we construct a curve γ slightly to the left of the maximal backward j-front ending at $\overline{y}$. Observe that every wave-front crossing γ from left to right must be of a family $i > j$. Moreover, we can

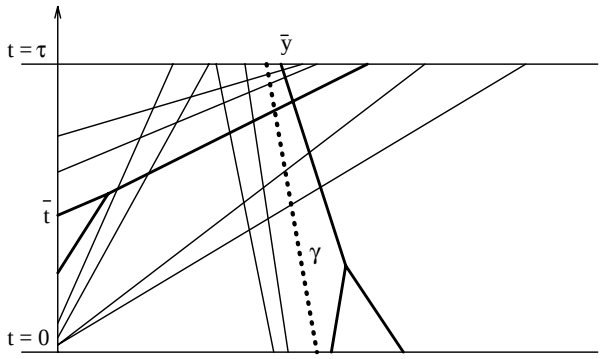


FIG. 6.

assume that no wave crosses γ from right to left. Applying Remark 4.2 to the region on the left of γ we obtain

$$\sum_{\alpha\in C(\gamma)} \xi_\alpha\,\sigma_\alpha + \sum_{x_\alpha(\tau)<\gamma(\tau)} \xi_\alpha(\tau)\,\sigma_\alpha(\tau) = \xi^0\,\sigma^0. \tag{5.11}$$

Since the waves crossing γ must belong to different families from the ones ending inside the interval $[0,\ \gamma(\tau)]$ (recall that interacting waves of the same family produce a single wave-front), (5.11) implies

$$\left|\sum_{\alpha\in C(\gamma)} \xi_\alpha\,\sigma_\alpha\right| \le C''\,|\xi^0\,\sigma^0| \tag{5.12}$$

for some constant C'' depending only on the system (1.1). We now again apply Remark 4.2 to the region on the right of γ, observing that the set of outgoing fronts, crossing the line $t=\tau$, contains the j-front at $\overline{y}$ plus other fronts on the right of $\overline{y}$ of families $i>j$. This yields

$$\xi_{\alpha^*}(\tau)\,\sigma_{\alpha^*}(\tau) + \sum_{x_\alpha(\tau)>\overline{y}} \xi_\alpha(\tau)\,\sigma_\alpha(\tau) = \sum_{\alpha\in C(\gamma)} \xi_\alpha\,\sigma_\alpha \tag{5.13}$$

which, in turn, implies

$$\left|\xi_{\alpha^*}(\tau)\,\sigma_{\alpha^*}(\tau)\right| = l_j(u^R)\cdot \sum_{\alpha\in C(\gamma)} \xi_\alpha\,\sigma_\alpha. \tag{5.14}$$

Observing that

$$l_j(u^L)\cdot \sum_{\alpha\in C(\gamma)} \xi_\alpha\,\sigma_\alpha = 0,$$

we again obtain an estimate of the form (5.10), and hence **(P3)** holds. The other cases are similar or easier.

We now complete the proof of Lemma 5.1. If we assume that only one single front is shifted leaving the boundary $x=0$, say, starting at time $t=\overline{t}$ (or starting at time $t=0$ and located at $x=\overline{x}$), it follows that at a fixed time $\tau>0$ the only fronts with nonzero shift rate can be the ones located inside the interval $[a_0,\ b_0] \doteq [0,\ \lambda^{\max}\cdot(\tau-\overline{t})]$

(or inside the interval $[a_0,\, b_0] \doteq [\max\{0, \overline{x}-\lambda^{\max}\cdot\tau\},\, \overline{x}+\lambda^{\max}\cdot\tau]$), where $\lambda^{\max}$ denotes the upper bound for the absolute value of all characteristic speeds in (2.7). Recalling the estimate (4.3)–(4.4) on the total variation, and using the properties (P1)–(P3), we thus have

$$\begin{aligned}\sum_{x_\alpha(\tau)>\delta} \big|\xi_\alpha(\tau)\,\sigma_\alpha(\tau)\big| &\leq n\,C_4|\xi^0\,\sigma^0| + C_5\big|\xi^0\,\sigma^0\big|\cdot \text{Tot.Var.}\big\{u(\tau)\,;\ [\delta,\ b_0]\big\}\\ &\leq C_3(1+N2^{-\nu})(1+\log(\tau/\delta))\big|\xi^0\,\sigma^0\big|\,,\end{aligned}\tag{5.15}$$

for a suitable constant C_3 depending only on the system (1.1), proving (5.2). Finally, we consider the case where all fronts in $\overline{u}$, $\tilde{u}$ are shifted. More precisely, let $\xi_\alpha(0)$ be the shift rate of the front located at $x_\alpha(0)$ ($t_\alpha(0)$, respectively), having amplitude $\sigma_\alpha(0)$. Call $\xi_\beta(\tau)$ the corresponding shift rate of the front of $u(\tau,\cdot)$ located at $x_\beta(\tau)$. Observing that the shift differential

$$\big(\xi_1(0),\dots\xi_M(0)\big)\mapsto\big(\xi_1(\tau),\dots\xi_{M'}(\tau)\big)$$

is a linear mapping, the estimate (5.2) follows easily from (5.15). □

In order to show that the trajectories of the flow map E_t that we shall construct in section 6 provide solutions with a strong $\mathbf{L}^1$ trace at the boundary $x=0$, we will make use of the following estimates on shift differentials of paths of approximate solutions along vertical segments of the domain Ω.

LEMMA 5.2. *Let $u(t,x)=E^\nu_t(\,\overline{u},\,\widetilde{u}\,)(x)$ be a front tracking solution containing at most N shocks. Assume that the fronts of $\overline{u}$ located at x_β (respectively, the fronts of $f(\widetilde{u})$ entering the interior of the domain Ω and starting at t_β) are shifted with shift rate ξ_β (respectively, $\widetilde{\xi}_\beta=\xi_\beta/\lambda_{k_\beta}$) and have amplitude σ_β (respectively, $\widetilde{\sigma}_\beta=\lambda_{k_\beta}\,\sigma_\beta$). Then there exists some constant C_6 depending only on the system* (1.1) *and on the domain K such that, for any $\tau_2>\tau_1>0$, and for every $0<\rho<(\lambda^{\min}/2)\,\tau_1$, denoting with $\widetilde{\xi}_\alpha(\rho)$, $\widetilde{\sigma}_\alpha(\rho)$ the time-shift rates and the time-sizes of the fronts in $f(u(\cdot,\rho\,))$ crossing the line $\{(t,\rho)\ ;\ t\geq 0\}$ at time $t_\alpha(\rho)$, there holds*

$$\sum_{t_\alpha(\rho)\in[\tau_1,\ \tau_2]}\big|\widetilde{\xi}_\alpha(\rho)\,\widetilde{\sigma}_\alpha(\rho)\big|\leq C_6(1+N2^{-\nu})(1+\log(\tau_2/\tau_1))\cdot\left(\sum_\beta\big|\xi_\beta\,\sigma_\beta\big|+\sum_\beta\big|\widetilde{\xi}_\beta\,\widetilde{\sigma}_\beta\big|\right).\tag{5.16}$$

Proof. We give here only a sketch of the proof, since it is quite similar to the one of Lemma 5.1. The estimates (P1)–(P3) can be recovered with minor modifications. As an example, we establish the estimate (P3) for a front belonging to a family $j\in\{1,\dots n-p\}$, say, (time)-located at $t_{\alpha^*}(\cdot)$, that starts at time $t=0$ and crosses the segment $\{(t,\rho);\ t\in[\tau_1,\ \tau_2]\}$ at $\overline{s}=t_{\alpha^*}(\rho)$. Assume that only a single front of (time)-size $\widetilde{\sigma}^0$ is shifted, leaving the boundary $x=0$ at time $t=\overline{t}$, say, of the family $k\in\{n-p+1,\dots,n\}$, with (time)-shift $\widetilde{\xi}^0$, and no other front of the boundary data $\widetilde{u}$ or of the initial data $\overline{u}$ is shifted. After performing the usual simplifications, we reduce to the situation illustrated in Figure 7. Consider the straight line $t=\overline{s}$ and a curve γ running slightly to the left of the maximal backward j-front passing through $(\,\overline{s},\rho)$. Applying the divergence theorem to the region on the left of γ, and using (4.1), we obtain

$$\widetilde{\xi}^0\,\widetilde{\sigma}^0=\xi^0\,\sigma^0=\sum_{\alpha\in C(\gamma)}\xi_\alpha\,\sigma_\alpha+\sum_{x_\alpha(\overline{s})<\gamma(\overline{s})}\xi_\alpha(\overline{s})\,\sigma_\alpha(\overline{s}).\tag{5.17}$$

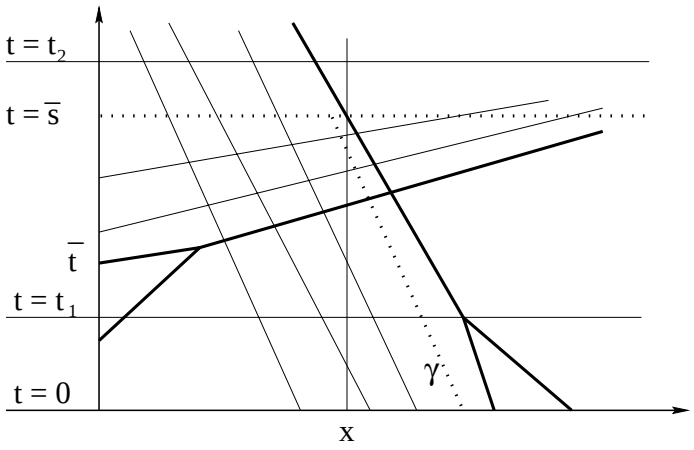


FIG. 7.

By linear independence of the vectors on the right-hand side of (5.17) we derive

$$\left| \sum_{\alpha\in C(\gamma)} \xi_\alpha\,\sigma_\alpha \right| \le C'''\,\big|\widetilde{\xi}^0\,\widetilde{\sigma}^0\big| \tag{5.18}$$

for some constant C''' depending only on the system (1.1). Next, we consider the region on the right of γ, where we compute

$$\sum_{\alpha\in C(\gamma)} \xi_\alpha\,\sigma_\alpha = \widetilde{\xi}_{\alpha^*}(\rho)\,\widetilde{\sigma}_{\alpha^*}(\rho) + \sum_{x_\alpha(\overline{s})>\rho} \xi_\alpha(\overline{s})\,\sigma_\alpha(\overline{s}). \tag{5.19}$$

From (5.18)–(5.19), and because of (2.1), (2.7), letting u^L, u^R denote as usual the left and right states across the jump at $(\overline{s},\rho)$, we obtain

$$\begin{aligned}\big|\widetilde{\xi}_{\alpha^*}(\rho)\,\widetilde{\sigma}_{\alpha^*}(\rho)\big| &= \left| l_j(u^R)\cdot \sum_{\alpha\in C(\gamma)} \xi_\alpha\,\sigma_\alpha \right| \\ &\le \big|l_j(u^R)-l_j(u^L)\big| \left| \sum_{\alpha\in C(\gamma)} \xi_\alpha\,\sigma_\alpha \right| \le \frac{C'^v}{\lambda^{\min}}\,\big|\widetilde{\sigma}_{\alpha^*}(\rho)\big|\big|\widetilde{\xi}^0\,\widetilde{\sigma}^0\big|\,,\end{aligned}$$

for some other constant C'^v depending only on the system (1.1), and we recover (P3).

Therefore, in the case where at the boundary $x=0$ only a single front is shifted, say, starting at time $t=\overline{t}$, observing that the only fronts of the last p characteristic families with nonzero shift rate along a fixed line $\{(t,\rho)\ ;\ t\ge 0\}$ can be the ones located inside the (time)-interval $[s_0,t_0]\doteq[\overline{t}+(\rho/\lambda^{\max}),\,\overline{t}+(\rho/\lambda^{\min})]$, one derives (5.17) relying on the properties (P1)–(P3) and using similar estimates on the total variation as the ones in (4.3)–(4.4). Namely, there will be some positive constant depending only on the system (1.1) that we may call C_2 as the one in (4.3)–(4.4) such that, for every $x>0$, and for any $t>s>0$, there holds

$$\text{Tot.Var.}\big\{w_k(\cdot,x);\ [s,\,t]\big\} \le 2C_2\cdot\log\frac{t}{s} + \|w_k\|_{L^\infty} + (N+1)2^{1-\nu} \tag{5.20}$$

for $k = 1, \dots, n-p$,

$$\text{(5.21)}\quad \text{Tot.Var.}\big\{w_k(\cdot, x);\ [s,\, t]\big\} \leq 2C_2\left\{\frac{t-s}{x} + \log\frac{t}{s}\right\} + \|w_k\|_{L^\infty} + (N+1)2^{1-\nu}$$

for $k = n-p+1, \dots, n$. The proof of the estimates (5.20)–(5.21) is entirely similar to the one of Lemma 4.4. With the same arguments we obtain (5.17) in the case where we assume that a single front is shifted at time $t = 0$ and located at $x = \overline{x}$, observing that, if $\rho > \overline{x}$, the fronts of the first p families with nonzero shift rate along the line $\{(t, \rho)\ ;\ t \geq 0\}$ are located inside the (time)-interval $[s_0, t_0] \doteq [0,\, \rho/\lambda^{\min}]$, while if $\rho < \overline{x}$, such fronts are the ones that interact on the left of $x = \rho$ with the front starting at $x = \overline{x}$, and hence their total strength is at most

$$\log\ \frac{\tau_2}{\tau_1 - \rho/\lambda^{\min}} \leq \log\ \frac{2\,\tau_2}{\tau_1}\,.$$

Finally, the general case where all fronts in $\overline{u}$ and in $\widetilde{u}$ are shifted is treated as in Lemma 5.1. □

Remark 5.1. If we perturb a front tracking solution $u(t, \cdot) \doteq E^\nu_t(\,\overline{u},\, \widetilde{u}\,)$ by shifting only the (time) locations of the jumps in the boundary data $\widetilde{u}$, with the same arguments of the proof of Lemma 5.2, one can show that the stability estimate (5.17) holds with a Lipschitz constant that is independent of τ_1, τ_2. Namely, in the same setting of Lemma 5.2, assuming that the fronts of $f(\widetilde{u})$, with (time)-size $\widetilde{\sigma}_\beta = \lambda_{k_\beta}\,\sigma_\beta$, are shifted with (time)-shift rate $\widetilde{\xi}_\beta = \xi_\beta/\lambda_{k_\beta}$, the following holds. There exists some constant (depending only on the system (1.1) and on the domain K) that we still call C_6 such that, for any fixed $\delta > 0$, and for every $0 < \rho < \lambda^{\min}\delta$, $\tau > \delta$, letting $\widetilde{\xi}_\alpha(\rho)$, $\widetilde{\sigma}_\alpha(\rho)$ be the time-shift rates and the time-sizes of the fronts in $f(u(\cdot, \rho\,))$ crossing the line $\{(t, \rho)\ ;\ t \geq 0\}$ at time $t_\alpha(\rho)$, there holds

$$\text{(5.22)}\qquad \sum_{t_\alpha(\rho)\in[\delta,\ \tau]} \big|\widetilde{\xi}_\alpha(\rho)\,\widetilde{\sigma}_\alpha(\rho)\big| \leq C_6(1 + N2^{-\nu}) \cdot \sum_\beta \big|\widetilde{\xi}_\beta\,\widetilde{\sigma}_\beta\big|.$$

6. Proof of Theorems 2.3 and 2.4.

6.1. Existence of the semigroup on domains of BV functions. In order to construct the semigroup described in Theorem 2.3, we shall first define an $\mathbf{L}^1$ continuous flow map E_t on every domain

$$\mathcal{D}_{\mathcal{M}} \doteq \Big\{\mathbf{p} \in \mathcal{D};\quad \text{Tot.Var.}\{\mathbf{p}\} \leq \mathcal{M}\Big\}, \qquad\qquad \mathcal{M} > 0,$$

obtained as a limit of the approximate flow maps E^ν_t constructed in section 3 on the domains $\mathcal{D}^\nu$. To this end, consider any two piecewise constant couples of initial and boundary data, say $\mathbf{p}_1,\ \mathbf{p}_2 \in \mathcal{D}_{\mathcal{M}} \cap \mathcal{D}^\nu$, and construct a pseudopolygonal path $\gamma^\nu_0 : \theta \mapsto \mathbf{fp}^\theta = (\,\overline{u}^\theta,\, f(\widetilde{u}^\theta))$ connecting $\mathbf{fp}_1$ with $\mathbf{fp}_2$ as described in section 3. All functions $(\,\overline{u}^\theta,\, \widetilde{u}^\theta\,)$ lie in $\mathcal{D}_{\mathcal{M}} \cap \mathcal{D}^\nu$ and have a uniformly bounded number of shocks, say $\leq N$. Call $u^\theta_\nu(t, \cdot) = E^\nu_t(\,\overline{u}^\theta,\, \widetilde{u}^\theta\,)$ the corresponding solution and consider the path $\gamma^\nu_t : \theta \mapsto \big(u^\theta_\nu(t, \cdot),\, f(\widetilde{u}^\theta)\big)$. Writing the length of this path in the form (3.13),

and using Lemma 5.1, for any fixed $\delta > 0$, and for every $t \geq \delta$, we obtain the estimate

$$\begin{aligned}
\big\|\rho^1_{\delta,+\infty}(\gamma^\nu_t)\big\|_{\mathbf{L}^1} &= \sum_{j=1}^m \int_{\theta_{j-1}}^{\theta_j} \sum_{\{\alpha\,:\,x^\theta_\alpha>\delta\}} \big|\Delta u^\theta_\nu(t,x^\theta_\alpha)\big|\left|\frac{\partial x^\theta_\alpha(t)}{\partial\theta}\right| d\theta \\
&\leq \sum_{j=1}^m \int_{\theta_{j-1}}^{\theta_j} C_3(1+N2^{-\nu})(1+\log(t/\delta)) \\
&\quad\cdot\left(\sum_\beta \big|\Delta u^\theta_\nu(0,x^\theta_\beta)\big|\left|\frac{\partial x^\theta_\beta(0)}{\partial\theta}\right| + \sum_{\{\beta'\,:\,t^\theta_{\beta'}<t\}} \big|\widetilde{\Delta} u^\theta_\nu(t^\theta_{\beta'},0)\big|\left|\frac{\partial t^\theta_{\beta'}(0)}{\partial\theta}\right|\right) d\theta \\
&\leq C_3(1+N2^{-\nu})(1+\log(t/\delta))\cdot\big\|\gamma^\nu_0\big\|_{0.0,t}\,, \qquad (6.1)
\end{aligned}$$

where $\rho^1_{\delta,+\infty}$ and $\|\cdot\|_{0,0,t}$ denote the restriction map and the seminorm introduced at (3.14)–(3.15). Observing that any function in $\mathcal{D}_\mathcal{M}\cap\mathcal{D}^\nu$ has at most $2^\nu\mathcal{M}$ jumps, from (6.1) we derive (3.17) with $L_{\mathcal{M},t} = C_3(1+\mathcal{M})(1+\log(t/\delta))$, which, in turn, because of (3.19)–(3.20), clearly implies (3.21).

Once we have established the uniform Lipschitz continuity of the maps $\mathbf{p}\mapsto E^\nu_t\,\mathbf{p}\restriction_{[\delta,+\infty[}$, on the domains $\mathcal{D}_\mathcal{M}\cap\mathcal{D}^\nu$, since the union $\cup_{\nu\geq1}\mathcal{D}_\mathcal{M}\cap\mathcal{D}^\nu$ is dense in $\mathcal{D}_\mathcal{M}$, we will define the map E_t on $\mathcal{D}_\mathcal{M}$ as the limit

$$E_t(\mathbf{p}) \doteq \mathbf{L}^1 - \lim_{\nu\to\infty} E^\nu_t(\mathbf{p}^\nu), \qquad \mathbf{p}^\nu\in\mathcal{D}_\mathcal{M}\cap\mathcal{D}^\nu,\ \mathbf{p}^\nu\to\mathbf{p}\text{ in }\mathbf{L}^1\,. \qquad (6.2)$$

In order to prove that the assignment (6.2) yields a well-defined map, since any sequence $E^\nu_t\mathbf{p}^\nu$ is uniformly bounded in $\mathbf{L}^\infty$, it is sufficient to show that, for every given $\mathbf{p}\in\mathcal{D}_\mathcal{M}$, and for any $\delta>0$, if $\mathbf{p}^\nu\in\mathcal{D}_\mathcal{M}\cap\mathcal{D}^\nu$ is any sequence that converges to $\mathbf{p}$ in $\mathbf{L}^1$, then the sequence $E^\nu_t\mathbf{p}^\nu\restriction_{[\delta,+\infty[}$ is Cauchy in $\mathbf{L}^1$. Indeed, for any $\mu>\nu$, using (3.21) (possibly with a different constant $L'_{\mathcal{M},t}$), we obtain

$$\begin{aligned}
\big\|E^\mu_t\mathbf{p}^\mu - E^\nu_t\mathbf{p}^\nu\big\|_{\mathbf{L}^1([\delta,+\infty[)} &\leq \big\|E^\mu_t\mathbf{p}^\mu - E^\mu_t\mathbf{p}^\nu\big\|_{\mathbf{L}^1([\delta,+\infty[)} + \big\|E^\mu_t\mathbf{p}^\nu - E^\nu_t\mathbf{p}^\nu\big\|_{\mathbf{L}^1([\delta,+\infty[)} \\
&\leq L'_{\mathcal{M},t}\cdot d_{0,0,t}(\mathbf{p}^\mu,\,\mathbf{p}^\nu) + d_{\delta,0,\infty}\big(S^\mu_t\mathbf{p}^\nu,\,S^\nu_t\mathbf{p}^\nu\big), \qquad (6.3)
\end{aligned}$$

where $d_{\delta,0,\infty}$ denotes the pseudometric defined as in (3.18). To estimate the second term in (6.3), we shall use the same type of error estimate established in [11, Theorem 2.9] for the distance between a Lipschitz continuous map and the trajectory of a Lipschitz continuous semigroup which can be restated as follows.

LEMMA 6.1. *Let $(B,\,d_B)$ be a metric space, let $d_{B'}$ be a pseudometric on B, and let $\mathcal{D}$ be a closed subset of B. Let $S:\mathcal{D}\times[0,\,T]\mapsto\mathcal{D}$ be a continuous semigroup and $\Gamma:[0,\,T]\mapsto\mathcal{D}$ a continuous map that satisfy*

$$d_{B'}\big(S_t\mathbf{p}_1,\,S_s\mathbf{p}_2\big)\leq L\cdot\big\{d_B(\mathbf{p}_1,\,\mathbf{p}_2)+|t-s|\big\}, \qquad (6.4)$$

$$d_B\big(\Gamma(t),\,\Gamma(s)\big)\leq L\cdot|t-s| \qquad (6.5)$$

for some constant $L>0$. Then, for any $\tau\in[0,\,T]$, one has the estimate

$$d_{B'}\big(\Gamma(\tau),\,S_\tau\Gamma(0)\big)\leq L\cdot\int_0^\tau\left\{\liminf_{h\to0+}\ \frac{d_{B'}\big(\Gamma(t+h),\,S_h\Gamma(t)\big)}{h}\right\}\,dt. \qquad (6.6)$$

Let B be the metric space $\mathbf{L}^1(\mathbb{R}^+, K) \times \mathbf{L}^1(\mathbb{R}^+, K)$ equipped with the usual $\mathbf{L}^1$ distance, and set

$$d_{B'} \doteq d_{\delta,0,\infty}, \qquad \mathcal{D} \doteq \mathcal{D}_{\mathcal{M}} \cap \mathcal{D}^\mu\,.$$

Observe that, if we let $S \doteq S^\mu$ be the approximate semigroup defined in (3.8), and $\Gamma : [0, T] \to \mathcal{D}_{\mathcal{M}} \cap \mathcal{D}^\nu \subset \mathcal{D}_{\mathcal{M}} \cap \mathcal{D}^\mu$ be the map $\Gamma(t) = S^\nu_t \mathbf{p}^\nu$, then the Lipschitz continuity (3.21) of $\mathbf{p} \mapsto E^\nu_t\, \mathbf{p}\,\restriction_{[\delta,+\infty[}$, together with the uniform bound on the total variation of $t \mapsto E^\nu_t \mathbf{p},\ \mathbf{p} \in \mathcal{D}_{\mathcal{M}} \cap \mathcal{D}^\nu,\ \ t \mapsto E^\mu_t \mathbf{p},\ \mathbf{p} \in \mathcal{D}_{\mathcal{M}} \cap \mathcal{D}^\mu$, clearly implies the estimates (6.4)–(6.5). Thus, we may apply Lemma 6.1 and, from (6.3), (6.6), we derive

$$\begin{aligned}\left\|E^\mu_t \mathbf{p}^\mu - E^\nu_t \mathbf{p}^\nu\right\|_{\mathbf{L}^1([\delta,+\infty[)} &\leq L'_{\mathcal{M},t} \cdot d_{0,0,t}(\mathbf{p}^\mu,\, \mathbf{p}^\nu)\\ &\quad + L \cdot \int_0^t \left\{ \liminf_{h\to 0+} \frac{d_{\delta,0,\infty}\big(S^\nu_{s+h}\, \mathbf{p}^\nu,\ S^\mu_h S^\nu_s \mathbf{p}^\nu\big)}{h} \right\} ds. \end{aligned} \tag{6.7}$$

With the same arguments in [6], letting $\mathbf{q} \doteq S^\nu_s \mathbf{p}^\nu$, we can now estimate the integrand in (6.7) by

$$\frac{1}{h}\, d_{\delta,0,\infty}\big(S^\nu_h\, \mathbf{q},\ S^\mu_h \mathbf{q}\big) = \frac{1}{h}\left\|E^\nu_h\, \mathbf{q} - E^\mu_h \mathbf{q}\right\|_{\mathbf{L}^1([\delta,+\infty[)} \leq C_7 \cdot 2^{-\nu} \mathcal{M} \tag{6.8}$$

for some constant $C_7 > 0$. Hence, (6.7) together with (6.8) yields

$$\left\|E^\mu_t \mathbf{p}^\mu - E^\nu_t \mathbf{p}^\nu\right\|_{\mathbf{L}^1([\delta,+\infty[)} \leq L'_{\mathcal{M},t} \cdot d_{0,0,t}(\mathbf{p}^\mu,\, \mathbf{p}^\nu) + LC_7 \mathcal{M} \cdot 2^{-\nu} t, \tag{6.9}$$

which clearly shows that $E^\nu_t \mathbf{p}^\nu\,\restriction_{[\delta,+\infty[}$ is a Cauchy sequence in the $\mathbf{L}^1$ norm and that this limit does not depend on the choice of the sequence $\mathbf{p}^\nu$. Thus, the map in (6.2) is well defined on every domain $\mathcal{D}_{\mathcal{M}}$ and, passing to the limit in (3.21), we obtain the estimate (2.25) for any couples of initial and boundary data $\mathbf{p}_i = (\,\overline{u}_i,\, \widetilde{u}_i\,) \in D_{\mathcal{M}},\ i = 1, 2$.

6.2. Extension of the semigroup to domains of $\mathbf{L}^\infty$ functions. To ensure the existence of the map E_t on the whole domain $\mathcal{D}$ of functions of possibly unbounded variation, we will now prove the estimate (3.22) for some constant $L''_t > 0$ independent of the total variation. To this purpose, consider any two couples $\mathbf{p}_1,\ \mathbf{p}_2 \in \mathcal{D}^\mu$, and construct as above a pseudopolygonal path $\gamma_0 : \theta \mapsto \mathbf{fp}^\theta = (\overline{u}^\theta,\, f(\widetilde{u}^\theta))$ taking values in $\mathcal{F}\mathcal{D}^\mu$ that connects $\mathbf{fp}_1$ with $\mathbf{fp}_2$ and has the following property. All functions $(\,\overline{u}^\theta,\, \widetilde{u}^\theta\,)$ have a uniformly bounded number of jumps and hence lie in some domain $\mathcal{D}_{\mathcal{M}},\ \mathcal{M} > 0$. Then, calling $u^\theta_\nu(t,\cdot) = E^\nu_t(\,\overline{u}^\theta,\, \widetilde{u}^\theta\,)$ the corresponding ν-approximate solution, since by (6.2) we have

$$E_t(\,\overline{u}^\theta,\, \widetilde{u}^\theta\,) = \lim_{\nu\to\infty} E^\nu_t(\,\overline{u}^\theta,\, \widetilde{u}^\theta\,) = \lim_{\nu\to\infty} u^\theta_\nu(t,\cdot), \tag{6.10}$$

in order to establish (3.22) we will show that the length of the path $\gamma^\nu_t : \theta \mapsto \big(u^\theta_\nu(t,\cdot),\, f(\widetilde{u}^\theta)\big)$ remains a bounded multiple of the length of γ_0, independent of ν. Indeed, for any fixed $\delta > 0$, and for every $\nu \geq \mu$, letting N be a uniform bound on the number of shocks in $(\,\overline{u}^\theta,\, \widetilde{u}^\theta\,)$, and using Lemma 5.1, we obtain by the same arguments in (6.1) the estimate

$$\left\|\rho^1_{\delta,+\infty}\big(\gamma^\nu_t\big)\right\|_{\mathbf{L}^1} \leq C_3(1 + N2^{-\nu})(1 + \log(t/\delta)) \cdot \|\gamma_0\|_{0,0,t} \qquad \forall\ t \geq \delta\,,$$

which, in turn, because of (3.16), (3.19)–(3.20), implies

(6.11)
$$\big\|E^\nu_t\mathbf{p}_1 - E^\nu_t\mathbf{p}_2\big\|_{\mathbf{L}^1([\delta,+\infty[)} \leq C_8(1+N2^{-\nu})(1+\log(t/\delta))\cdot d_{0,0,t}(\mathbf{fp}_1,\,\mathbf{fp}_2) \qquad \forall\, t\geq\delta$$

for some other constant $C_8>0$. Letting $\nu\to\infty$ in (6.12), because of (6.10) we obtain (3.22) for all $\mathbf{p}_1,\ \mathbf{p}_2\in\mathcal{D}^\mu$. Since the domains $\mathcal{D}^\mu,\ \mu\geq 1$, are dense in $\mathcal{D}$, relying on (3.22) we can now extend the map E_t by continuity to the whole domain $\mathcal{D}$ setting

$$E_t(\mathbf{p}) \doteq \mathbf{L}^1\!-\lim_{\mu\to\infty} E_t(\mathbf{p}^\mu), \qquad\qquad \mathbf{p}^\mu\in\mathcal{D}^\mu,\ \mathbf{p}^\mu\to\mathbf{p}\ \text{in}\ \mathbf{L}^1\,. \tag{6.12}$$

Clearly, the map in (6.12) preserves the property (3.22), proving (2.25). Moreover, any trajectory $t\mapsto E_t(\mathbf{p})$, being the limit of front tracking approximations, provides by standard arguments [10, 11] a weak solution to problem (1.1)–(1.2).

6.3. Stability estimates in space. Towards a proof of the existence of a strong $\mathbf{L}^1$ trace of $f(u(t,x))\doteq f\big(E_t\mathbf{p}(x)\big)$ at the boundary $x=0$, we shall first establish the stability estimate (3.24) for the map $\mathbf{p}\mapsto f\big(E_{(\cdot)}\mathbf{p}(x)\big)$. Fix $\tau_2>\tau_1>0$, and observe that, because of (6.2), for every given $\mathbf{p}\in\mathcal{D}_{\mathcal{M}}$, the sequence $E^\nu_{(\cdot)}\mathbf{p}(\cdot)$ converges to $E_{(\cdot)}\mathbf{p}(\cdot)$ in $\mathbf{L}^1\big([0,\,\tau_2]\times\mathbb{R}^+;\,K\big)$. Hence, relying also on the continuity of the maps $x\mapsto E^\nu_{(\cdot)}\mathbf{p}(x),\ x\mapsto E_{(\cdot)}\mathbf{p}(x)$, we deduce that

(6.13)
$$f\big(E_{(\cdot)}(\mathbf{p})(x)\big)\restriction_{[\tau_1,\,\tau_2]} = \mathbf{L}^1-\lim_{\nu\to\infty} f\big(E^\nu_{(\cdot)}(\mathbf{p})(x)\big)\restriction_{[\tau_1,\,\tau_2]} \qquad \forall\, x\in[0,\,(\lambda^{\min}/2)\,\tau_1].$$

Therefore we may proceed as in the proof of (3.22) to establish the estimate (3.24). Given any pair of couples $\mathbf{p}_1,\ \mathbf{p}_2\in\mathcal{D}^\mu$, we construct a pseudopolygonal path $\gamma_0:\theta\mapsto\mathbf{fp}^\theta=(\,\overline{u}^\theta,\,f(\widetilde{u}^\theta))$ taking values in $\mathcal{FD}^\mu$, and connecting $\mathbf{fp}_1$ with $\mathbf{fp}_2$, so that all functions $(\,\overline{u}^\theta,\,\widetilde{u}^\theta)$ have a uniformly bounded number of shocks $\leq N$ and lie in some domain $\mathcal{D}_{\mathcal{M}},\ \mathcal{M}>0$. Then, for every $\nu\geq\mu$, calling $u^\theta_\nu(t,\cdot)\doteq E^\nu_t(\,\overline{u}^\theta,\,\widetilde{u}^\theta\,)(\cdot)$ the corresponding ν-approximate solution, we consider the pseudopolygonal path

$$\gamma^\nu_x:\theta\mapsto\big(\,\overline{u}^\theta,\,f\big(u^\theta_\nu(\cdot,\,x)\big)\big) \tag{6.14}$$

with values in $\mathcal{FD}^\nu$. Let $\rho^i_{\tau_1,\tau_2}$ and $\|\cdot\|_{0,0,\tau_2}$ denote the restriction map and the seminorm defined as in (3.14)–(3.15). Then, using Lemma 5.2, we compute as in (6.1)

$$\begin{aligned}
\big\|\rho^2_{\tau_1,\tau_2}(\gamma^\nu_x)\big\|_{\mathbf{L}^1} &= \sum_{j=1}^m\int_{\theta_{j-1}}^{\theta_j}\sum_{\{\alpha\,:\,t^\theta_\alpha\in[\tau_1,\tau_2]\}}\big|\widetilde{\Delta}u^\theta_\nu(t^\theta_\alpha,x)\big|\left|\frac{\partial t^\theta_\alpha(x)}{\partial\theta}\right|d\theta\\
&\leq \sum_{j=1}^m\int_{\theta_{j-1}}^{\theta_j} C_6(1+N2^{-\nu})(1+\log(\tau_2/\tau_1))\\
&\qquad\cdot\left(\sum_\beta\big|\Delta u^\theta_\nu(0,x^\theta_\beta)\big|\left|\frac{\partial x^\theta_\beta(0)}{\partial\theta}\right|+\sum_{\{\beta'\,:\,t^\theta_{\beta'}<\tau_2\}}\big|\widetilde{\Delta}u^\theta_\nu(t^\theta_{\beta'},0)\big|\left|\frac{\partial t^\theta_{\beta'}(0)}{\partial\theta}\right|\right)d\theta\\
&\leq C_6(1+N2^{-\nu})(1+\log(\tau_2/\tau_1))\cdot\|\gamma_0\|_{0,0.\tau_2}
\end{aligned} \tag{6.15}$$

for every $x\in[0,\,(\lambda^{\min}/2)\,\tau_1]$. Observe that the $\mathbf{L}^1$ length of the path γ^ν_x satisfies

$$\big\|\gamma^\nu_x\big\|_{0,\tau_1,\tau_2} = \big\|\rho^1_{0,+\infty}(\gamma^\nu_0)\big\|_{\mathbf{L}^1}+\big\|\rho^2_{\tau_1,\tau_2}(\gamma^\nu_x)\big\|_{\mathbf{L}^1}\,, \tag{6.16}$$

$$\big\|f\big(E^\nu_{(\cdot)}\mathbf{p}_1(x)\big)-f\big(E^\nu_{(\cdot)}\mathbf{p}_2(x)\big)\big\|_{\mathbf{L}^1([\tau_1,\,\tau_1])}\leq C_9\cdot\|\gamma^\nu_x\|_{0,\tau_1,\tau_2}\,, \tag{6.17}$$

for every $x \in [0,\, (\lambda^{\min}/2)\,\tau_1]$, and for some constant $C_9 > 0$. Hence, recalling also (3.19), we deduce from (6.15) the estimate

$$\begin{aligned}\big\|f\big(E^\nu_{(\cdot)}\mathbf{p}_1(x)\big)-f\big(E^\nu_{(\cdot)}\mathbf{p}_2(x)\big)\big\|_{\mathbf{L}^1([\tau_1,\,\tau_1])}\\ \leq C_{10}(1+N2^{-\nu})(1+\log(\tau_2/\tau_1))\cdot d_{0,0,\tau_2}(\mathbf{fp}_1,\,\mathbf{fp}_2)\end{aligned}\tag{6.18}$$

for some other constant $C_{10} > 0$. Letting $\nu \to \infty$ in (6.18), thanks to (6.14) we obtain (3.24) with $L''' = C_{10}(1 + \log(\tau_2/\tau_1))$. By continuity, and relying on the density of the domains $\mathcal{D}^\mu,\ \mu \geq 1$ in $\mathcal{D}$, we then extend the estimate (3.24) to any pair $\mathbf{p}_1,\ \mathbf{p}_2$ in $\mathcal{D}$.

Remark 6.1. If we fix a piecewise constant initial data $\overline{u} \in \mathbf{L}^1(\mathbb{R}^+, K^\mu)$, and for any given pair of piecewise constant boundary data $\widetilde{u},\, \widetilde{v} \in \mathbf{L}^1(\mathbb{R}^+, K^\nu)$, $\nu \geq \mu$, we construct a pseudopolygonal path $\gamma_0 : \theta \mapsto \mathbf{fp}^\theta = (\,\overline{u},\, f(\widetilde{u}^\theta))$, taking values in $\mathcal{FD}^\mu$, and connecting $\mathbf{fp}_1 \doteq (\,\overline{u},\, f(\widetilde{u}))$ with $\mathbf{fp}_2 \doteq (\,\overline{u},\, f(\widetilde{v}))$, with the same arguments above, and relying on Remark 5.1, we derive the same type of estimate as (3.24) with a Lipschitz constant that is independent on $\tau_1,\, \tau_2$. Thus, by continuity, and by the density of the domains $\mathcal{D}^\mu,\ \mu \geq 1$ in $\mathcal{D}$, we deduce that there exists some constant C', depending only on the system (1.1), so that for any fixed $\tau > \delta > 0$, and for all $(\,\overline{u},\, \widetilde{u}\,), (\,\overline{u},\, \widetilde{v}\,) \in \mathcal{D}$, there holds

$$\begin{aligned}\big\|E_{(\cdot)}\big(\overline{u},\, \widetilde{u}\,\big)(x) - E_{(\cdot)}\big(\overline{u},\, \widetilde{v}\,\big)(x)\big\|_{\mathbf{L}^1([\delta,\,\tau])} \leq C'\cdot\big\|\widetilde{u}-\widetilde{v}\big\|_{\mathbf{L}^1([0,\,\tau])}\\ \forall\ x \in [0,\, \lambda^{\min}\,\delta]\,.\end{aligned}\tag{6.19}$$

6.4. Oleinik-type estimates. Concerning the entropy admissibility conditions (2.15)–(2.16) on the decay of the positive waves, consider a couple of initial data and boundary conditions $(\,\overline{u},\, \widetilde{u}\,) \in \mathcal{D}$ and fix any interval $[a, b]$. Thanks to (3.21)–(3.22), we can now approximate the weak solution constructed as above, $u(t,\cdot) = E_t(\,\overline{u},\, \widetilde{u}\,)$, with a sequence of front tracking solutions $u^\nu(t,\cdot) = E^\nu_t(\,\overline{u}^\nu,\, \widetilde{u}^\nu\,)$, choosing initial and boundary data $(\,\overline{u}^\nu,\, \widetilde{u}^\nu\,) \in \mathcal{D}^\nu$ having a number of shocks $N_\nu \leq \nu$. By (4.5)–(4.6), the total number of positive wave-fronts in $u^\nu(\tau,\cdot) = E^\nu_\tau(\,\overline{u}^\nu,\, \widetilde{u}^\nu\,)$ on $[a, b]$ satisfies

$$\text{Pos.Var.}\big\{w^\nu_k(\tau,\cdot);\ [a,\ b]\big\} \leq C_2\cdot\frac{b-a}{\tau} + (N_\nu+1)2^{1-\nu}\tag{6.20}$$

for $k = 1, \ldots, n-p$,

$$\text{Pos.Var.}\big\{w^\nu_k(\tau,\cdot);\ [a,\ b]\big\} \leq C_2\cdot\left\{\frac{b-a}{\tau} + \log\left(\frac{b}{a}\right)\right\} + (N_\nu+1)2^{1-\nu}\,,\tag{6.21}$$

for $k = n-p+1, \ldots, n$, where $w^\nu_k \doteq w_k(u^\nu)$ denotes as usual the kth Riemann coordinate of u^ν. Letting $\nu \to \infty$ in (6.20)–(6.21), by the lower semicontinuity of the total variation we obtain (2.15)–(2.16). The estimates (2.17)–(2.18) on the decay of the positive variation of $w_k(\cdot\,, x)$ can be established in the entirely similar way relying on the corresponding estimates for $w^\nu_k(\cdot, x)$ which, in turn, are obtained with the same type of arguments used to prove the ones in (5.20)–(5.21).

6.5. Boundary conditions. Let $u(t,x) \doteq E_t\mathbf{p}(x),\ \mathbf{p} = (\,\overline{u},\, \widetilde{u}\,) \in D$, be the weak solution defined at (6.12), and consider a sequence $\mathbf{p}^\nu = (\,\overline{u}^\nu,\, \widetilde{u}^\nu\,) \in \mathcal{D}^\nu$ converging to $\mathbf{p}$ in $\mathbf{L}^1$ as $\nu \to \infty$. Call $u^\nu(t,x) \doteq E_t\mathbf{p}^\nu(x)$ the corresponding solution. Since every $\mathbf{p}^\nu$ is piecewise constant and lies in some domain $D_{\mathcal{M}_\nu}$, one can easily

verify that any function $u^\nu(t,x)$, $\nu \geq 1$, has bounded total variation and pointwise satisfies the boundary condition (2.20); i.e., there holds

$$\lim_{x\to 0+} w_j\big(u^\nu(t,x)\big) = w_j\big(\widetilde{u}^\nu(t)\big) \quad \text{for a.e.} \quad t\geq 0, \quad j = n-p+1,\dots,n. \tag{6.22}$$

Now, fix $\tau_2 > \tau_1 > 0$. By (3.24) and because of the invertibility property of the flux function, f, there will be some constant $C_{11} = C_{11}(\tau_1,\tau_2) > 0$ (depending only on τ_1,τ_2) such that

$$\big\|w_j\big(u^\nu(\cdot,x)\big) - w_j\big(u(\cdot,x)\big)\big\|_{\mathbf{L}^1([\tau_1,\ \tau_2])} \leq C_{11}\cdot d_{0,0,\tau_2}(\mathbf{p}^\nu,\ \mathbf{p}) \tag{6.23}$$

for all $x\in[0,\ (\lambda^{\min}/2)\,\tau_1]$, $\nu\geq 1$. Then, (6.22), (6.23) together imply that, for any $j = n-p+1,\dots,n$, the functions $w_j\big(u(\cdot,x)\big)$, $w_j\big(f(u(\cdot,x))\big)$ have a strong limit as $x\to 0$ and

$$\lim_{x\to 0+}\int_{\tau_1}^{\tau_2} \big|w_j\big(u(t,x)\big) - w_j\big(\widetilde{u}(t)\big)\big|dt = 0\,, \tag{6.24}$$

thus showing that $u(t,x)$ fulfills the boundary condition (2.20). On the other hand, because of the Oleinik-type conditions (2.15) on the decay of the positive waves, also $w_j\big(u(\cdot,x)\big)$, $j=1,\dots,n-p$, have a strong limit as $x\to 0$, which completes the proof of the existence of the strong $\mathbf{L}^1$ trace of $u(t,x)$ at $x=0$, and hence concludes the proof of Theorem 2.3.

6.6. Uniqueness. Let u be an entropy weak solution to (1.1)–(1.3) on the region $\Omega_T \doteq [0,\,T]\times\mathbb{R}^+$ in accordance with Definition 2.2, and assume that conditions (i)–(iii) stated in Theorem 2.4 hold. Let $\lambda^{\max}$ be the upper bound for the absolute value of all characteristic speeds at (2.7), and fix $R > \lambda^{\max}\cdot T$, $0<\delta<(R-\lambda^{\max}\cdot T)/2$. Observe that, because of the entropy conditions (2.15)–(2.16) on the decay of the positive waves, for every fixed $0<s\leq\delta$, the restrictions of $u(t,\cdot)$ to the intervals $J_{\delta,R}(t)\doteq[2\delta,\,R-\lambda^{\max}\cdot t]$, $s\leq t\leq T$, have uniformly bounded total variation. Thus, the same type of uniqueness results in [13] yield

$$u(t,\cdot) = E_{t-s}\big(u(s,\,s+\cdot),\ u(s+\cdot,\,s)\big)(-s+\cdot) \qquad \text{restricted to} \quad J_{\delta,R}(t) \tag{6.25}$$

for every $0<s\leq\delta\leq t\leq T$. Moreover, by the definition of $J_{\delta,R}(t)$, the domain of dependence of a solution to (1.1)–(1.3) along the segment $\{(t,x)\ ;\ x\in J_{\delta,R}(t)\}$ is contained in the set $\{(s,x)\in\mathbb{R}^2\ ;\ 0\leq s\leq t,\quad 0\leq x\leq R\}$. Hence, recalling Remark 2.5, we can restate the Lipschitz estimate (2.25) provided by Theorem 2.3 as

$$\begin{aligned}
&\int_{2\delta-s}^{R-(\lambda^{\max}\cdot t+s)} \Big|E_{t-s}\big(u(s,s+\cdot),u(s+\cdot,s)\big)(x)\\
&\qquad\qquad -E_{t-s}\big(E_s(\overline{u},\widetilde{u})(s+\cdot),\ E_{(s+\cdot)}(\overline{u},\widetilde{u})(s)\big)(x)\Big|dx\\
&\leq C(1+\log(t/\delta))\cdot\Big\{\int_\delta^R \Big|u(s,\,x) - E_s(\overline{u},\widetilde{u})(x)\Big|dx\\
&\qquad\qquad + \sum_{j=n-p+1}^{n}\int_s^t \Big|w_j\big(u(\sigma,\,s)\big) - w_j\big(E_\sigma(\,\overline{u},\,\widetilde{u}\,)(s)\big)\Big|d\sigma\Big\},
\end{aligned} \tag{6.26}$$

which, because of (6.25), yields

$$\begin{aligned}&\int_{J_{\delta,R}(t)} \Big|u(t,x) - E_t(\,\overline{u},\,\widetilde{u}\,)(x)\Big|dx\\ &\le C(1+\log(t/\delta))\cdot\Big\{\int_0^R \Big|u(s,\,x) - E_s(\overline{u},\widetilde{u})(x)\Big|dx\\ &\qquad + \sum_{j=n-p+1}^{n}\int_0^t \Big|w_j\big(u(\sigma,\,s\big) - w_j\big(E_\sigma(\,\overline{u},\,\widetilde{u}\,)(s)\big)\Big|d\sigma\Big\}\end{aligned} \tag{6.27}$$

for every $0 < s \le \delta \le t \le T$. On the other hand, the continuity in $\mathbf{L}^1{}_{loc}$ of the map $t \mapsto E_t(\,\overline{u},\,\widetilde{u}\,)$ at $t = 0$ (see Remark 2.4), and the existence of a strong $\mathbf{L}^1$ trace of $E_t(\,\overline{u},\,\widetilde{u}\,)(x)$ at the boundary $x = 0$ (guaranteed by Theorem 2.3), together with (2.19), imply

$$\begin{aligned}&\lim_{s\to 0^+}\int_0^R \Big|E_s(\overline{u},\widetilde{u})(x) - \overline{u}(x)\Big|dx = 0\,,\\ &\lim_{s\to 0^+}\int_0^t \Big|w_j\big(E_\sigma(\,\overline{u},\,\widetilde{u}\,)(s) - w_j\big(\widetilde{u}(\sigma)\big)\Big|\,d\sigma = 0\,.\end{aligned} \tag{6.28}$$

Thus, taking the essential limit of the right-hand side of (6.27) as $s \to 0^+$, using (6.28), and relying on (2.31)–(2.33), we obtain

$$\int_{2\delta}^{R-\lambda^{\max}\cdot t} \big|u(t,x) - E_t(\,\overline{u},\,\widetilde{u}\,)(x)\big|dx = 0 \qquad \forall\, t\in[0,T]\,. \tag{6.29}$$

Since $\delta \in]0,\, R - \lambda^{\max}\cdot T)/2[\,$, and $R > \lambda^{\max}\cdot T$ were arbitrary, this concludes the proof of Theorem 2.4.

7. Proof of Theorem 2.6. We give here only the proof of the compactness of the attainable sets $\mathcal{A}(T,\mathcal{U})$, $T > 0$, in connection with the sets of admissible boundary controls $\mathcal{U}$ defined in (2.39), the procedure to establish the compactness of $\mathcal{A}(\,\overline{x},\mathcal{U})$, $\overline{x} > 0$, being entirely similar.

Fix $T > 0$. Given $\overline{u} \in \mathbf{L}^1(\mathbb{R}^+, K)$, consider a sequence of boundary data $\{\widetilde{u}^\nu\}_{\nu\ge 0} \subset \mathcal{U}$, and let $u^\nu(t,x) \doteq E_t(\,\overline{u},\,\widetilde{u}^\nu\,)(x)$ be the corresponding solutions. Observe that all solutions $u^\nu(t,x)$, $\nu \ge 0$, are uniformly bounded since they take values in the compact set K. Moreover, thanks to the Oleinik-type estimates (2.15)–(2.16) on the time decay of the positive waves, for every fixed $0 < a < b$, and $0 < \tau \le T$, there exist constants $C' = C'(a,b,\tau) > 0$, $C'' = C''(a,b,\tau) > 0$ such that

$$\text{Tot.Var.}\big\{u^\nu(t,\cdot)\ ;\ [a,\,b]\big\} \le C' \qquad \forall\, t\in[\tau,\,T], \qquad \forall\,\nu\ge 0\,, \tag{7.1}$$

$$\int_a^b |u^\nu(t,x) - u^\nu(s,x)|\,dx \le C''|t-s| \qquad \forall\, t,s\in[\tau,\,T], \qquad \forall\,\nu\ge 0\,. \tag{7.2}$$

Hence, applying Helly's theorem, we deduce that there exists a subsequence $\{u^{\nu_j}\}_{j\ge 0}$ so that $\{u^{\nu_j}(t,\cdot)\restriction_{[a,\,b]}\}_{j\ge 0}$ converges in $\mathbf{L}^1([a,b],\,K)$ to some function $u_{a,b,\tau}(t,\cdot)$ for any $t \in [\tau,\,T]$. Therefore, by considering sequences of real number $a_k \to 0+$, $b_k \to \infty$, $\tau_k \to 0+$, and by using a diagonal procedure, we construct a subsequence $\{u^{\nu'}(t,\cdot)\}_{\nu'\ge 0}$ that converges in $\mathbf{L}^1{}_{loc}(\mathbb{R}^+,K)$ to some function $u(t,\cdot)$ for any $t \in [0,\,T]$. We claim that there exists a boundary data $\widetilde{u} \in \mathcal{U}$ such that

$$u(t,\cdot) = E_t(\,\overline{u},\,\widetilde{u}\,) \qquad \forall\, t\in[0,\,T]\,. \tag{7.3}$$

Notice that, by construction, the map $(t,x) \to \big(u(t,\cdot),\, u(\cdot,x)\big)$ takes values within the domain $\mathcal{D}_T$ defined in (2.30). Moreover, the estimate (7.2) implies the continuity of $u : [0,\, T] \times \mathbb{R}^+ \mapsto U$ as a function from $]0,\, T]$ into $\mathbf{L}^1{}_{loc}(\mathbb{R}^+)$. Hence, thanks to Theorem 2.4 and Lemma 2.5, in order to prove the claim it will be sufficient to show the following:

(1) There exists a boundary data $\widetilde{u} \in \mathcal{U}$ so that $u(t,x)$ is an entropy weak solution to (1.1)–(1.3) on the region $[0,\, T] \times \mathbb{R}^+$, in the sense of Definition 2.2.
(2) Given any boundary entropy pair $\big(\alpha(u,v),\, \beta(u,v)\big)$ for (1.1), there is a constant $M > 0$ (depending only on $(\alpha,\, \beta)$ and on the domain K) for which $u(t,x)$ satisfies the corresponding distributional entropy inequality (2.34).

Towards a proof of (1) observe that, because of (2.26), all fluxes $f(u^\nu)$, $\nu \geq 0$, admit a strong $\mathbf{L}^1$ trace Ψ^ν at $x = 0$, whose essential range is contained in the compact set $f(K)$. Hence, the sequence $\{\Psi^\nu\}_{\nu \geq 0}$ is weak* relatively compact in $\mathbf{L}^\infty(\mathbb{R}^+)$ and, by possibly taking a subsequence, we have

$$\Psi^\nu \,\overset{*}{\rightharpoonup}\, \Psi \qquad \text{in} \quad \mathbf{L}^\infty(\mathbb{R}^+) \tag{7.4}$$

for some function $\Psi \in \mathbf{L}^\infty(\mathbb{R}^+, \mathbb{R}^n)$. Moreover, by Theorem 2.3, every u^ν is a distributional solution of the corresponding initial-boundary value problem on $[0,\, T] \times \mathbb{R}^+$; i.e., there holds

$$\begin{aligned} &\int_0^T \int_0^{+\infty} \Big\{ u^\nu(t,x) \cdot \phi_t(t,x) + f(u^\nu(t,x)) \cdot \phi_x(t,x) \Big\}\, dx\, dt \\ &\qquad + \int_0^{+\infty} \overline{u}(x) \cdot \phi(0,x)\, dx + \int_0^T \Psi^\nu(t) \cdot \phi(t,0)\, dt = 0 \end{aligned} \tag{7.5}$$

for any test function $\phi \in \mathcal{C}^1_c$ with compact support contained in the set $]-\infty,\, T\,[\times\mathbb{R}$. Therefore, passing to the limit as $\nu \to \infty$ in (7.5), we get

$$\begin{aligned} &\int_0^T \int_0^{+\infty} \Big\{ u(t,x) \cdot \phi_t(t,x) + f(u(t,x)) \cdot \phi_x(t,x) \Big\}\, dx\, dt \\ &\qquad + \int_0^{+\infty} \overline{u}(x) \cdot \phi(0,x)\, dx + \int_0^T \Psi(t) \cdot \phi(t,0)\, dt = 0\,. \end{aligned} \tag{7.6}$$

By considering, in particular, test functions $\phi \in \mathcal{C}^1_c$ with compact support contained in the set $]-\infty,\, T\,[\times]0,\, \infty[$, from (7.6) we deduce that $u(t,x)$ is a distributional solution of the Cauchy problem (1.1)–(1.2) on the region $[0,\, T] \times \mathbb{R}^+$, as required by Definition 2.2(i). On the other hand, given any $\mathcal{C}^1$ function $\alpha = \alpha(t)$, writing (7.6) for test functions $\phi^\varepsilon(t,x) = \alpha(t) \cdot \beta^\varepsilon(t,x)$, $\beta^\varepsilon \in \mathcal{C}^1_c$, supported on the semistrips $]0,\, T[\times]-\infty,\, \varepsilon[$, $\varepsilon > 0$, shrinking to the region $]0,\, T[\times]-\infty,\, 0]$ around the boundary $]0,\, T[\times\{0\}$, and such that $\phi^\varepsilon(t,0) = \alpha(t)$, we obtain

$$\lim_{x \to 0^+} \int_0^T f(u(t,x)) \cdot \alpha(t)\, dt = \int_0^T \Psi(t) \cdot \alpha(t)\, dt\,, \tag{7.7}$$

thus proving (2.13).

Now observe that, by Remark 2.2, the definition (2.39) of the set $\mathcal{U}$ of admissible boundary data implies

$$w_j\big(f^{-1}(\Psi^\nu(t))\big) \in [c_j,\, d_j] \qquad \text{for a.e.} \quad t \geq 0, \qquad j = n-p+1, \ldots, n. \tag{7.8}$$

Hence, for every flux trace Ψ^ν, $\nu \geq 0$, one has

(7.9)
$$\Psi^\nu(t) \in \mathcal{G} \doteq \Big\{ f(u) \ ; \ w_j(u) \in [c_j,\, d_j] \qquad \forall\ j = n-p+1, \dots, n \Big\} \qquad \text{for a.e.} \quad t \geq 0.$$

Since, by the properties of the Riemann invariants, the set $\mathcal{G}$ is closed and convex, it follows that the weak limit of Ψ^ν satisfies

$$\Psi(t) \in \mathcal{G} \qquad \text{for a.e.} \quad t \geq 0. \tag{7.10}$$

Therefore, if we consider the boundary data $\widetilde{u}$ defined in Riemann coordinates by

$$w_j(\widetilde{u}(t)) \doteq \begin{cases} \overline{\gamma}_j & \text{if} \quad j \leq n-p, \\ \\ w_j\big(f^{-1}(\Psi(t))\big) & \text{if} \quad j > n-p, \end{cases} \qquad \forall\ t \geq 0\,, \tag{7.11}$$

for some constant values

$$\overline{\gamma}_j \in \begin{cases} [c_j,\, d_j] & \text{if} \quad j \in J, \\ \\ [a_j,\, b_j] & \text{otherwise} \end{cases} \qquad j \leq n-p\,, \tag{7.12}$$

(J denoting the set of indices in the definition (2.39) of $\mathcal{U}$, and a_j, b_j being the constants in the definition (2.5) of the set K), we clearly have $\widetilde{u} \in \mathcal{U}$, and, by Remark 2.2, $u(t,x)$ satisfies the boundary condition (2.14) of Definition 2.2(ii). To complete the proof of (1) observe that the Oleinik-type conditions of Definition 2.2(iii) can be recovered by the lower semicontinuity of the total variation, since the map $u(t,x)$ is obtained as the $\mathbf{L}^1{}_{loc}$ limit of a sequence of maps satisfying (2.15)–(2.18).

Finally, regarding (2) observe that every solution $t \mapsto u^\nu(t,\cdot)$, being a trajectory of the flow map E_t, is obtained as the limit of front tracking approximations, and hence by standard arguments satisfies the distributional entropy inequality (2.34) associated with any boundary entropy pair for (1.1). Clearly, this property is preserved by the $\mathbf{L}^1{}_{loc}$ limit u of the sequence u^ν. This concludes the proof of Theorem 2.6.

8. Appendix. We show here that, if the system (1.1) is of Temple class, the two sets of admissible boundary values $\mathcal{V}^{\mathcal{E}ntr}(\widetilde{u})$, $\mathcal{V}(\widetilde{u})$ defined in (2.12), (2.22) coincide. Indeed, it was already shown in [8] that, if $u \in \mathcal{V}^{\mathcal{E}ntr}(\widetilde{u})$, and η is an entropy for the system (1.1) that is differentiable in $\widetilde{u}$, with $D\eta(\widetilde{u}) = 0$, while q is the corresponding flux associated with η, then u satisfies the entropy inequality that appears in the definition (2.22), since $q(u) \leq q(\widetilde{u})$. Therefore, we deduce that u satisfies all the inequalities in (2.22) associated with entropies η that are differentiable in $\widetilde{u}$, since in this case one can write $\eta = \eta_1 + \eta_2$, with

$$\eta_1(u) \doteq \eta(\widetilde{u}) + D\eta(\widetilde{u}) \cdot (u - \widetilde{u}), \qquad \eta_2(u) \doteq \eta(u) - \eta_1(u)\,, \tag{8.1}$$

where $D\eta_2(\widetilde{u}) = 0$, while η_1 is an affine entropy which trivially verifies the inequality in (2.22). By density it follows that u satisfies all the inequalities in (2.22) associated with any (continuous) convex entropy η, proving

$$\mathcal{V}(\widetilde{u}) \subseteq \mathcal{V}^{\mathcal{E}ntr}(\widetilde{u})\,. \tag{8.2}$$

We will show below that the converse inclusion also holds.

LEMMA 8.1. *Let* (1.1) *be a system of Temple class, and let* K *be a set as in* (2.5). *Assume that the strict hyperbolicity condition* (SH1) *is verified and that, for*

some index $p \in \{1, \dots, n\}$, there holds (2.6). *Then, letting $\mathcal{V}(\widetilde{u})$, $\mathcal{V}^{\mathcal{E}ntr}(\widetilde{u})$ be the sets of admissible boundary values defined, respectively, in* (2.12), *and in* (2.22), *one has*

$$\mathcal{V}^{\mathcal{E}ntr}(\widetilde{u}) \subseteq \mathcal{V}(\widetilde{u}) \qquad \forall\, \widetilde{u} \in K\,. \tag{8.3}$$

Proof. First observe that, by Remark 2.2, if $w = (w_1, \dots, w_n)$ is a system of Riemann coordinates for (1.1), then the inclusion (8.3) is verified if and only if, for every $\widetilde{u} \in K$, there holds

$$u \in \mathcal{V}^{\mathcal{E}ntr}(\widetilde{u}) \qquad \Longrightarrow \qquad w_j(u) = w_j(\widetilde{u}) \qquad \forall\, j = n-p+1, \dots, n\,. \tag{8.4}$$

Next, fix $\widetilde{u} \in K$, and, for every $j > n-p$, consider the following Kruzkow-type entropy-entropy flux pair (see, e.g., [29, Chapter 13]):

$$\begin{aligned} \eta_j(u) &= \big|l_j(\widetilde{u}) \cdot (u - \widetilde{u})\big|\,, \\ q_j(u) &= l_j(\widetilde{u}) \cdot (f(u) - f(\widetilde{u}))\ \mathrm{sgn}\big(l_j(\widetilde{u}) \cdot (u - \widetilde{u})\big)\,, \end{aligned} \tag{8.5}$$

where $l_j(u)$ denotes the left eigenvector of the Jacobian matrix $DF(u)$, normalized as in (2.1). We will show that, if we set

$$\mathcal{E}(\eta,\, q;\, \zeta,\, u) \doteq q(u) - q(\widetilde{u}) - \zeta \cdot \big(f(u) - f(\widetilde{u})\big), \qquad \zeta \in \partial\eta(\widetilde{u}) \tag{8.6}$$

($\partial\eta(\widetilde{u})$ denoting the subdifferential of η at $\widetilde{u}$), then, for any $j > n-p$, there holds

$$\mathcal{E}(\eta_j,\, q_j;\, \zeta,\, u) \le 0 \qquad \forall\, \zeta \in \partial\eta(\widetilde{u}) \qquad \Longrightarrow \qquad l_j(\widetilde{u}) \cdot (u - \widetilde{u}) = 0\,, \tag{8.7}$$

which proves (8.4) since $l_j(\widetilde{u}) \cdot (u - \widetilde{u}) = 0$ is the equation of the hyperplane $\{u \in U\ ;\ w_j(u) = w_j(\widetilde{u})\}$. Observe that $\partial\eta(\widetilde{u}) = \{\gamma\, l_j(\widetilde{u})\ ;\ \gamma \in \partial|\cdot| = [-1,\, 1]\}$, and recall that for Temple class systems there exists a smooth, matrix-valued map $M : \mathbb{R}^n \times \mathbb{R}^n \to M_{n\times n}(\mathbb{R})$ with the following properties (see [20]):

(i) There holds

$$\begin{aligned} f(u) - f(v) &= M(u, v) \cdot (u - v) \qquad \forall\, u,\, v\,, \\ M(u, u) &= Df(u) \qquad \forall\, u. \end{aligned} \tag{8.8}$$

(ii) $M(u, v)$ and $Df(v)$ have the same (left and right) eigenvectors.

Multipling both terms of the first equality in (8.8) on the left by $l_j(v)$, we obtain

$$l_j(v) \cdot (f(u) - f(v)) = \lambda_j(u, v)\, l_j(v) \cdot (u - v) \qquad \forall\, u,\, v\,, \tag{8.9}$$

where $\lambda_j(u, v)$ denotes the ith eigenvalue of $M(u, v)$. Then, using (8.9), by a direct computation we find

$$\begin{aligned} &\mathcal{E}\big(\eta_j,\, q_j;\, \zeta,\, u\big) \\ &\quad = l_j(\widetilde{u}) \cdot (f(u) - f(\widetilde{u}))\ \mathrm{sgn}\big(l_j(\widetilde{u}) \cdot (u - \widetilde{u})\big) - \gamma\, l_j(\widetilde{u}) \cdot (f(u) - f(\widetilde{u})) \\ &\quad = \lambda_j(u, \widetilde{u})\Big(\big|l_j(\widetilde{u}) \cdot (u - \widetilde{u})\big| - \gamma\, l_j(\widetilde{u}) \cdot (u - \widetilde{u})\Big) \qquad \forall\, \gamma \in [-1,\, 1]\,. \end{aligned} \tag{8.10}$$

Since, by (2.6), one has $\lambda_j(u, \widetilde{u}) > 0$ for any $j > n-p$, from (8.10) we deduce (8.7), thus concluding the proof. ∎

Acknowledgment. The authors would like to thank Dr. Stefano Bianchini for many useful discussions.

REFERENCES

[1] D. Amadori, *Initial-boundary value problems for nonlinear systems of conservation laws*, NoDEA Nonlinear Differential Equations Appl., 4 (1997), pp. 1–42.

[2] D. Amadori and R. M. Colombo, *Continuous dependence for* 2×2 *conservation laws with boundary*, J. Differential Equations, 138 (1997), pp. 229–266.

[3] D. Amadori and R. M. Colombo, *Characterization of viscosity solutions for conservation laws with boundary*, Rend. Sem. Mat. Univ. Padova, 99 (1998), pp. 219–245.

[4] F. Ancona and A. Marson, *On the attainable set for scalar nonlinear conservation laws with boundary control*, SIAM J. Control Optim., 36 (1998), pp. 290–312.

[5] F. Ancona and A. Marson, *Scalar non-linear conservation laws with integrable boundary data*, Nonlinear Anal., 35 (1999), pp. 687–710.

[6] P. Baiti and A. Bressan, *The semigroup generated by a Temple class system with large data*, Differential Integral Equations, 10 (1997), pp. 401–418.

[7] C. Bardos, A. Y. Leroux, and J. C. Nedelec, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[8] A. Benabdallah and D. Serre, *Problèmes aux limites pour des systèmes hyperboliques non linéaires de deux équations à une dimension d'espace*, C. R. Acad. Sci. Paris Sér. I Math., 305 (1987), pp. 677–680.

[9] S. Bianchini, *Stability of* L^∞ *solutions for hyperbolic systems with coinciding shocks and rarefactions*, SIAM J. Math. Anal., 33 (2001), pp. 959–981.

[10] A. Bressan, *Global solutions of systems of conservation laws by wave-front tracking*, J. Math. Anal. Appl., 170 (1992), pp. 414–432.

[11] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One-Dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000.

[12] A. Bressan and R. M. Colombo, *The semigroup generated by* 2×2 *conservation laws*, Arch. Rational Mech. Anal., 133 (1995), pp. 1–75.

[13] A. Bressan and P. Goatin, *Oleinik type estimates and uniqueness for* $n \times n$ *conservation laws*, J. Differential Equations, 156 (1999), pp. 26–49.

[14] A. Bressan and P. Goatin, *Stability of* $\mathbf{L}^\infty$ *solutions of Temple class systems*, Differential Integral Equations, 13 (2000), pp. 1503–1528.

[15] G.-Q. Chen and H. Frid, *Divergence-measure fields and hyperbolic conservation laws*, Arch. Ration. Mech. Anal., 147 (1999), pp. 89–118.

[16] G.-Q. Chen and H. Frid, *Vanishing viscosity limit for initial-boundary value problems for conservation laws*, Contemp. Math. 238, AMS, Providence, RI, 1999, pp. 35–51.

[17] F. Dubois and P. G. LeFloch, *Boundary conditions for non-linear hyperbolic systems of conservation laws*, J. Differential Equations, 71 (1988), pp. 93–122.

[18] B. Dubroca and G. Gallice, *Résultats d'existence et d'unicité du probléme mixte pour des systémes hyperboliques de lois de conservation monodimensionels*, Comm. Partial Differential Equations, 15 (1990), pp. 59–80.

[19] J. Goodman, *Initial Boundary Value Problems for Hyperbolic Systems of Conservation Laws*, Ph.D. thesis, Stanford University, Stanford, CA, 1982.

[20] A. Heibig, *Existence and uniqueness of solutions for some hyperbolic systems of conservation laws*, Arch. Rational Mech. Anal., 126 (1994), pp. 79–101.

[21] K. T. Joseph and P. G. LeFloch, *Boundary layers in weak solutions of hyperbolic conservation laws*, Arch. Ration. Mech. Anal., 147 (1999), pp. 47–88.

[22] B. L. Keyfitz, *Solutions with shocks*, Comm. Pure Appl. Math., 24 (1971), pp. 125–132.

[23] H. O. Kreiss, *Initial-boundary value problems for hyperbolic systems*, Comm. Pure Appl. Math., 23 (1970), pp. 277–298.

[24] P. G. LeFloch, *Explicit formula for scalar non-linear conservation laws with boundary condition*, Math. Methods Appl. Sci., 10 (1988), pp. 265–287.

[25] T. P. Liu, *Initial-boundary value problem for gas dynamics*, Arch. Rational Mech. Anal., 64 (1977), pp. 137–168.

[26] T. P. Liu, *The free piston problem for gas dynamics*, J. Differential Equations, 30 (1978), pp. 175–191.

[27] F. Otto, *Initial-boundary value problem for a scalar conservation law*, C. R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 729–734.

[28] M. Sablé-Tougeron, *Méthode de Glimm et probléme mixte*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 423–443.

[29] D. Serre, *Systemes de Lois de Conservation*, Diderot Editeur, Paris, 1996.

[30] B. Temple, *Systems of conservation laws with invariant submanifolds*, Trans. Amer. Math. Soc., 280 (1983), pp. 781–795.

SIAM J. MATH. ANAL.
Vol. 34, No. 1, pp. 64–86

A REFINED GLOBAL WELL-POSEDNESS RESULT FOR SCHRÖDINGER EQUATIONS WITH DERIVATIVE*

J. COLLIANDER†, M. KEEL‡, G. STAFFILANI§, H. TAKAOKA¶, AND T. TAO‖

Abstract. In this paper we prove that the one-dimensional Schrödinger equation with derivative in the nonlinear term is globally well-posed in H^s for $s > \frac{1}{2}$ for data small in L^2. To understand the strength of this result one should recall that for $s < \frac{1}{2}$ the Cauchy problem is ill-posed, in the sense that uniform continuity with respect to the initial data fails. The result follows from the method of almost conserved energies, an evolution of the "I-method" used by the same authors to obtain global well-posedness for $s > \frac{2}{3}$. The same argument can be used to prove that any quintic nonlinear defocusing Schrödinger equation on the line is globally well-posed for large data in H^s for $s > \frac{1}{2}$.

Key words. almost conserved energies, global well-posedness, Schrödinger equation with derivative

AMS subject classifications. 35Q53, 42B35, 37K10

PII. S0036141001394541

1. Introduction.

In this paper, using the method of almost conserved energies, we establish a sharp result on global well-posedness for the derivative nonlinear Schrödinger IVP

$$\begin{cases} i\partial_t u + \partial_x^2 u = i\lambda\partial_x(|u|^2 u), \\ u(x,0) = u_0(x), \qquad\qquad x \in \mathbb{R},\ t \in \mathbb{R}, \end{cases} \tag{1}$$

where $\lambda \in \mathbb{R}$.

The first result of this kind was obtained in the context of the KdV and the modified KdV (mKdV) IVPs [11], also using almost conserved energies. Below we will discuss in more detail the "almost conservation method" and its relationship with the "I-method" which was applied to (1) in [9] (see also [10, 11, 20, 21]).

From the point of view of physics the equation in (1) is a model for the propagation of circularly polarized Alfvén waves in magnetized plasma with a constant magnetic field [25, 26, 29].

*Received by the editors August 28, 2001; accepted for publication (in revised form) March 1, 2002; published electronically August 15, 2002.
http://www.siam.org/journals/sima/34-1/39454.html

†Department of Mathematics, University of Toronto, Toronto, ON M5S 3G3 Canada (colliand@math.toronto.edu). This author was supported in part by NSF grant DMS-0100595.

‡Department of Mathematics, University of Minnesota, 127 Vincent Hall, 206 Church St. S.E., Minneapolis, MN 55455 (keel@math.umn.edu). This author was supported in part by NSF grant DMS-9801558.

§Brown University and Stanford University. Current address: Department of Mathematics, Box 1917, Brown University, Providence, RI 02912 (gigliola@math.brown.edu). This author was supported in part by NSF grant DMS-0100375 and grants from Hewlett-Packard and the Sloan Foundation.

¶Department of Mathematics, Hokkaido University, Sapporo, 060-0810, Japan (takaoka@math.sci.hokudai.ac.jp). This author was supported in part by JSPS grant 13740087.

‖Department of Mathematics, University of California, 405 Hilgard Ave., Los Angeles, CA 90095 (tao@math.ucla.edu). This author is a Clay Prize Fellow and was supported in part by a grant from the Packard Foundation.

It is natural to impose the smallness condition

$$\|u_0\|_{L^2} < \sqrt{\frac{2\pi}{|\lambda|}} \tag{2}$$

on the initial data, as this will force the energy to be positive via the sharp Gagliardo–Nirenberg inequality [36]. Note that the L^2 norm is conserved by the evolution. In this paper, we prove the following global well-posedness result.

THEOREM 1.1. *The Cauchy problem* (1) *is globally well-posed in* H^s *for* $s > \frac{1}{2}$, *assuming the smallness condition* (2).

We present here once again [9] a summary of the well-posedness story for (1). Scattering and well-posedness for this Cauchy problem has been studied by many authors [14, 15, 16, 17, 18, 19, 27, 28, 30, 34, 35]. The best local well-posedness result is due to Takaoka [30], where a gauge transformation and the Fourier restriction method are used to obtain local well-posedness in H^s, $s \geq \frac{1}{2}$. In [31], Takaoka showed this result is sharp in the sense that, when $s < \frac{1}{2}$, the nonlinear evolution $u(0) \mapsto u(t)$, thought of as a map from H^s to H^s for some fixed t, fails to be C^3 or even uniformly C^0 in this topology, even when t is arbitrarily close to zero and the H^s norm of the data is small (see also Bourgain [5] and Biagioni–Linares [2]). Therefore, we see that Theorem 1.1 is sharp, in the sense described above, except for the endpoint.

In [27], global well-posedness is obtained for (1) in H^1 assuming the smallness condition (2). The argument there is based on two gauge transformations performed in order to remove the derivative in the nonlinear term and the conservation of the Hamiltonian. This was improved by Takaoka [31], who proved global well-posedness in H^s for $s > \frac{32}{33}$ assuming (2). His method of proof is based on the idea of Bourgain [4, 6] of estimating separately the evolution of low frequencies and of high frequencies of the initial data. In [9], we used the "I-method" to further push the Sobolev exponent for global well-posedness down to $s > \frac{2}{3}$. The main idea of the "I-method" consists of defining a modified H^s norm permitting us to capture some nonlinear cancellations in frequency space during the evolution (1). These cancellations allow us to prove that the modified $H^s(\mathbb{R})$ norm is nearly conserved in time, and an iteration of the local result proves global well-posedness provided $s > \frac{2}{3}$. In this paper, an algorithmic procedure, first developed in the KdV context [11], is applied to better capture the cancellations in frequency space. Successive applications of the algorithm generate higher-order-in-u but lower-order-in-scaling corrections to the modified H^s norm. After one application of our algorithm, we show that the modified H^s norm *with* the generated correction terms changes less in time than the modifed H^s norm itself, so the first application of the algorithm produces an *almost conserved energy.* The improvement obtained allows us to iterate the local result and prove global well-posedness in $H^s(\mathbb{R})$ provided $s > \frac{1}{2}$. In principle, the algorithm may itself be iterated to generate a sequence of almost conserved energies giving further insight into the dynamical properties of (1). The endpoint $s = \frac{1}{2}$ is not obtained here. We speculate, however, that a further refinement of the "almost conservation method" could be a possible way to approach this question.

We conclude this section with the following remark.

Remark 1.2. Consider the one-dimensional quintic nonlinear Schrödinger

$$i\partial_t u = \partial_{xx} u + iau\bar{u}\partial_x u + ibu^2\partial_x \bar{u} + cu^3\bar{u}^2, \tag{3}$$

where a, b, and c are fixed real numbers. If $(a+b)(3a-5b)/48+c/3 < 0$ the equation in (3) is defocusing and, as was remarked in [9], the techniques used to prove Theorem 1.1

apply here too, and one can prove global well-posedness for initial data in $H^s, s > \frac{1}{2}$. Moreover, if $a = b = 0$, we expect our method to give global well-posedness[1] even below $s = 1/2$.

We should point out that Clarkson and Cosgrove [8] (see also [1]) proved that (3) fails the Painlevé test for complete integrability when

$$c \neq \frac{1}{4}b(2b - a).$$

In particular, this shows that our techniques, which do not depend on a, b, c, do not rely on complete integrability.

2. Notation and known facts. To prove Theorem 1.1 we may assume $\frac{1}{2} < s \leq \frac{2}{3}$, since for $s > \frac{2}{3}$ the result is contained in [27, 31] and [9]. Henceforth $\frac{1}{2} < s \leq \frac{2}{3}$ shall be fixed. Also, by rescaling u, we may assume $\lambda = 1$.

We use C to denote various constants depending on s; if C depends on other quantities as well, this will be indicated by explicit subscripting; e.g., $C_{\|u_0\|_2}$ will depend on both s and $\|u_0\|_2$. We use $A \lesssim B$ to denote an estimate of the form $A \leq CB$, and $A \sim B$ for $cB \leq A \leq CB$, where c and C are absolute constants. We also use $A \ll B$ if $A \leq \epsilon B$, where ϵ is a very small absolute constant. We use $a+$ and $a-$ to denote expressions of the form $a + \varepsilon$ and $a - \varepsilon$, where $0 < \varepsilon \ll 1$ depends only on s.

We use $\|f\|_p$ to denote the $L^p(\mathbb{R})$ norm and $L^q_t L^r_x$ to denote the mixed norm

$$\|f\|_{L^q_t L^r_x} := \left(\int \|f(t)\|_r^q \, dt \right)^{1/q}$$

with the usual modifications when $q = \infty$.

We define the spatial Fourier transform of $f(x)$ by

$$\mathcal{F}(f)(\xi) := \hat{f}(\xi) := \int_{\mathbb{R}} e^{-ix\xi} f(x) \, dx$$

and the spacetime Fourier transform $u(t, x)$ by

$$\widetilde{\mathcal{F}}(u)(\tau, \xi) := \tilde{u}(\tau, \xi) := \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-i(x\xi + t\tau)} u(t, x) \, dtdx.$$

Note that the derivative ∂_x is conjugated to multiplication by $i\xi$ by the Fourier transform.

We shall also define D_x to be the Fourier multiplier with symbol $\langle \xi \rangle := 1 + |\xi|$. We can then define the Sobolev norms H^s by

$$\|f\|_{H^s} := \|D_x^s f\|_2 = \|\langle \xi \rangle^s \hat{f}\|_{L^2_\xi}.$$

We also define the spaces $X^{s,b}(\mathbb{R}\times\mathbb{R})$ (first introduced in the context of the Schrödinger equation in [3]; see also [22, 23]) on $\mathbb{R} \times \mathbb{R}$ by

$$\|u\|_{X^{s,b}(\mathbb{R}\times\mathbb{R})} := \|\langle \xi \rangle^s \langle \tau - |\xi|^2 \rangle^b \hat{u}(\xi, \tau)\|_{L^2_\tau L^2_\xi}.$$

[1] Recall that in this case the initial value problem is locally well-posed in H^s for $s \geq 0$; see [7] and [33].

We often abbreviate $\|u\|_{s,b}$ for $\|u\|_{X^{s,b}(\mathbb{R}\times\mathbb{R})}$. For any time interval I, we define the restricted spaces $X^{s,b}(I\times\mathbb{R})$ by

$$\|u\|_{X^{s,b}(I\times\mathbb{R})} := \inf\{\|U\|_{s,b} : U|_{I\times\mathbb{R}} = u\}.$$

We shall take advantage of the Strichartz estimate (see, e.g., [3])

$$\|u\|_{L_t^6 L_x^6} \lesssim \|u\|_{0,\frac{1}{2}+}, \tag{4}$$

which interpolates with the trivial estimate

$$\|u\|_{L_t^2 L_x^2} \lesssim \|u\|_{0,0}, \tag{5}$$

to give

$$\|u\|_{L_t^p L_x^p} \lesssim \|u\|_{0,\alpha(p)} \tag{6}$$

for any $p\in[2,6]$ and $\alpha(p) = \frac{(3+)(p-2)}{4p}$. We also use

$$\|u\|_{L_t^\infty L_x^2} \lesssim \|u\|_{0,\frac{1}{2}+}, \tag{7}$$

which together with Sobolev embedding gives

$$\|u\|_{L_t^\infty L_x^\infty} \lesssim \|u\|_{\frac{1}{2}+,\frac{1}{2}+}. \tag{8}$$

The next lemma introduces two more estimates that are probably less known than the standard Strichartz estimates.

LEMMA 2.1. *For any $b > \frac{1}{2}$ and any function u for which the right-hand side is well defined, we have*

$$\|D_x^{\frac{1}{2}} u\|_{L_x^\infty L_t^2} \lesssim \|u\|_{X^{0,b}} \tag{9}$$

(smoothing effect estimate).

For any $s > \frac{1}{2}$ and $\rho \geq \frac{1}{4}$ we have

$$\|u\|_{L_x^2 L_t^\infty} \lesssim \|u\|_{X^{s,b}}, \tag{10}$$

$$\|u\|_{L_x^4 L_t^\infty} \lesssim \|u\|_{X^{\rho,b}} \tag{11}$$

(maximal function estimates).

Proof. The estimates (9), (10), and (11) come from estimating the solution $S(t)u_0$ of the linear one-dimensional Schrödinger IVP in the norm appearing in the left-hand side and from a standard argument of summation along parabolic curves; see, for example, the expository paper [13]. The smoothing effect and maximal function estimates for $S(t)u_0$ can be found, for example, in [24]. □

We also have the following improved Strichartz estimate (cf. Lemma 7.1 in [9]; see also [4, 28, 32]).

LEMMA 2.2. *For any Schwartz functions u, v with Fourier support in $|\xi| \sim R$, $|\xi| \ll R$, respectively, we have that*

$$\|uv\|_{L_t^2 L_x^2} = \|u\bar{v}\|_{L_t^2 L_x^2} \lesssim R^{-1/2}\|u\|_{0,1/2+}\|v\|_{0,1/2+}.$$

In our arguments we shall be using the trivial embedding

$$\|u\|_{s_1,b_1} \lesssim \|u\|_{s_2,b_2} \quad \text{whenever } s_1 \leq s_2, b_1 \leq b_2$$

so frequently that we will not mention this embedding explicitly.

We now give some useful notation for multilinear expressions. If $n \geq 2$ is an even integer, we define a *(spatial) n-multiplier* to be any function $M_n(\xi_1, \dots, \xi_n)$ on the hyperplane

$$\Gamma_n := \{(\xi_1, \dots, \xi_n) \in \mathbb{R}^n : \xi_1 + \cdots + \xi_n = 0\},$$

which we endow with the standard measure $\delta(\xi_1 + \cdots + \xi_n)$, where δ is the Dirac delta.

If M_n is an n-multiplier and $f_1, \dots, f_n$ are functions on $\mathbb{R}$, we define the *n-linear functional* $\Lambda_n(M_n; f_1, \dots, f_n)$ by

$$\Lambda_n(M_n; f_1, \dots, f_n) := \int_{\Gamma_n} M_n(\xi_1, \dots, \xi_n) \prod_{j=1}^{n} \hat{f}_j(\xi_j).$$

We adopt the notation

$$\Lambda_n(M_n; f) := \Lambda_n(M_n; f, \bar{f}, f, \bar{f}, \dots, f, \bar{f}).$$

Observe that $\Lambda_n(M_n; f)$ is invariant under permutations of the even ξ_j indices or of the odd ξ_j indices.

If M_n is a multiplier of order n, $1 \leq j \leq n$ is an index, and $k \geq 1$ is an even integer, we define the *elongation* $\mathbf{X}_j^k(M_n)$ of M_n to be the multiplier of order $n + k$ given by

$$\mathbf{X}_j^k(M_n)(\xi_1, \dots, \xi_{n+k}) := M_n(\xi_1, \dots, \xi_{j-1}, \xi_j + \ldots + \xi_{j+k}, \xi_{j+k+1}, \dots, \xi_{n+k}).$$

In other words, $\mathbf{X}_j^k$ is the multiplier obtained by replacing ξ_j by $\xi_j + \cdots + \xi_{j+k}$ and advancing all the indices after ξ_j accordingly.

We shall often write ξ_{ij} for $\xi_i + \xi_j$, ξ_{ijk} for $\xi_i + \xi_j + \xi_k$, etc. We also write ξ_{i-j} for $\xi_i - \xi_j$, ξ_{ij-klm} for $\xi_{ij} - \xi_{klm}$, etc. Also, if $m(\xi)$ is a function defined in the frequency space, we use the notation $m(\xi_i) = m_i$, $m(\xi_{ij-k}) = m_{ij-k}$, etc.

In this paper we often use two very elementary tools: the mean value theorem (MVT) and the double mean value theorem (DMVT). While recalling the statement of the MVT will be an embarrassment, we think that doing so for the DMVT is a necessity to avoid later confusion.

LEMMA 2.3 (DMVT). *Assume $f \in C^2(\mathbb{R})$ and that $\max(|\eta|, |\lambda|) \ll |\xi|$; then*

$$|f(\xi + \eta + \lambda) - f(\xi + \eta) - f(\xi + \lambda) + f(\xi)| \lesssim |f''(\theta)||\eta||\lambda|,$$

where $|\theta| \sim |\xi|$.

3. The gauge transformation, energy, and the almost conservation laws. In this section we summarize the main results presented in section 3 and 4 of [9]. Whatever is here simply stated and recalled is fully explained or proved in those sections.

We start by applying the gauge transform used in [27] in order to improve the derivative nonlinearity present in (1).

DEFINITION 3.1. *We define the nonlinear map $\mathcal{G} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ by*

$$\mathcal{G}f(x) := e^{-i\int_{-\infty}^{x} |f(y)|^2 dy} f(x).$$

The inverse transform $\mathcal{G}^{-1}f$ is then given by

$$\mathcal{G}^{-1}f(x) := e^{i\int_{-\infty}^{x}|f(y)|^2dy}f(x).$$

This transform is a bicontinuous map from H^s to itself for any $s \in [0,1]$.

Set $w_0 := \mathcal{G}u_0$, and $w(t) := \mathcal{G}u(t)$ for all times t. A straightforward calculation shows that the IVP (1) transforms into

$$\begin{cases} i\partial_t w + \partial_x^2 w = -iw^2\partial_x\bar{w} - \frac{1}{2}|w|^4 w, \\ w(x,0) = w_0(x), \qquad\qquad x \in \mathbb{R},\ t \in \mathbb{R}. \end{cases} \tag{12}$$

Also, the smallness condition (2) becomes

$$\|w_0\|_{L^2} < \sqrt{2\pi}. \tag{13}$$

By the bicontinuity we thus see that global well-posedness of (1) in H^s is equivalent to that of (12). From [27, 30, 31], we know that both Cauchy problems are locally well-posed in $H^s, s \geq \frac{1}{2}$, and globally well-posed in H^1 assuming (13). By standard limiting arguments, we thus see that Theorem 1.1 will follow if we can show the following.

PROPOSITION 3.2. *Let w be a global H^1 solution to* (12) *obeying* (13). *Then for any $T > 0$ and $s > \frac{1}{2}$ we have*

$$\sup_{0\leq t\leq T}\|w(t)\|_{H^s} \lesssim C_{(\|w_0\|_{H^s},T)},$$

where the right-hand side does not depend on the H^1 norm of w.

We now pass to the considerations on the energy associated with solutions of (12).

DEFINITION 3.3. *If $f \in H^1(\mathbb{R})$, we define the energy $E(f)$ by*

$$E(f) := \int \partial_x f \partial_x \overline{f}\ dx - \frac{1}{2}\mathrm{Im}\int f\overline{f}f\partial_x\overline{f}\ dx.$$

By the Gagliardo–Nirenberg inequality we have

$$\|\partial_x f\|_2 \leq C_{\|f\|_2}E(f)^{1/2} \tag{14}$$

for any $f \in H^1$ such that $\|f\|_2 < \sqrt{2\pi}$.

By Plancherel, we write $E(f)$ using the Λ notation and Fourier transform properties as

$$E(f) = -\Lambda_2(\xi_1\xi_2; f) - \frac{1}{2}\mathrm{Im}\Lambda_4(i\xi_4; f). \tag{15}$$

Expanding out the second term using $\mathrm{Im}(z) = (z - \bar{z})/2i$, and using symmetry, we may rewrite this as

$$E(f) = -\Lambda_2(\xi_1\xi_2; f) + \frac{1}{8}\Lambda_4(\xi_{13-24}; f). \tag{16}$$

One can use the same notation to rewrite the L^2 norm as

$$\|w(t)\|_2^2 = \Lambda_2(1; w(t)).$$

LEMMA 3.4 (see [27]). *If w is an H^1 solution to* (12) *for $t \in [0,T]$, then we have*

$$\|w(t)\|_2 = \|w_0\|_2$$

and

$$E(w(t)) = E(w_0)$$

for all $t \in [0,T]$.

In [9] this lemma was proved using the following general proposition (cf. [9]).

PROPOSITION 3.5. *Let $n \geq 2$ be an even integer, let M_n be a multiplier of order n, and let w be a solution of* (12). *Then*

$$\begin{aligned}\partial_t \Lambda_n(M_n; w(t)) &= i\Lambda_n\left(M_n \sum_{j=1}^{n}(-1)^j \xi_j^2; w(t)\right)\\ &\quad - i\Lambda_{n+2}\left(\sum_{j=1}^{n} \mathbf{X}_j^2(M_n)\xi_{j+1}; w(t)\right)\\ &\quad + \frac{i}{2}\Lambda_{n+4}\left(\sum_{j=1}^{n}(-1)^{j-1}\mathbf{X}_j^4(M_n); w(t)\right).\end{aligned} \tag{17}$$

We summarize below the idea we used to prove Proposition 3.2 for $s > \frac{2}{3}$ in [9]. Because we do not want to use the H^1 norm of w, we cannot directly use the energy $E(w(t))$ defined above. So we introduced a substitute notion of "energy" that could be defined for a less regular solution and that had a very slow increment in time. In frequency space consider an even C^∞ monotone multiplier $m(\xi)$ taking values in $[0, 1]$ such that

$$m(\xi) := \begin{cases} 1 & \text{if } |\xi| < N, \\ \left(\frac{|\xi|}{N}\right)^{s-1} & \text{if } |\xi| > 2N. \end{cases} \tag{18}$$

Define the multiplier operator $I : H^s \longrightarrow H^1$ such that $\widehat{Iw}(\xi) := m(\xi)\widehat{w}(\xi)$. This operator is smoothing of order $1 - s$; indeed one has

$$\|u\|_{s_0,b_0} \lesssim \|Iu\|_{s_0+1-s,b_0} \lesssim N^{1-s}\|u\|_{s_0,b_0} \tag{19}$$

for any $s_0, b_0 \in \mathbb{R}$. Our substitute energy was defined by

$$E_N(w) := E(Iw).$$

Note that this energy makes sense even if w is only in H^s. In general, the energy $E_N(w(t))$ is not conserved in time, but we showed that the increment was very small in terms of N.

To proceed with the improvement of the "I-method," let us consider a symmetric multiplier $m(\xi)^2$ and let I be the multiplier operator associated with it. Then we write

$$E^1(w) := E(Iw).$$

Clearly, if m is the multiplier in (18), then

$$E^1(w) = E_N(w),$$

so we can think about $E^1(w)$ as the first generation of a family of modified energies. In this paper we introduce the second generation in detail, but formally the method can be used to define an infinite family of modified energies. We write

$$E^2(w) = -\Lambda_2(m_1\xi_1 m_2\xi_2, w) + \frac{1}{2}\Lambda_4\left(M_4(\xi_1, \xi_2, \xi_3, \xi_4), w\right), \tag{20}$$

[2]This eventually will be taken to be exactly the multiplier in (18).

where M_4 will be determined later. Assume now that w is a solution of (12). Because w is fixed we drop it from the definition of E^2. We are interested in the increment of this second generation of energies, and hence we compute $\frac{d}{dt}E^2$. Differentiating $\Lambda_2(m_1\xi_1 m_2\xi_2)$ using Proposition 3.5, using the identity $\xi_1+\cdots+\xi_n=0$ and symmetrizing, we have

$$\begin{aligned}\frac{d}{dt}\Lambda_2(m_1\xi_1 m_2\xi_2) =& -i\Lambda_2(m_1\xi_1 m_2\xi_2(\xi_1^2-\xi_2^2)) - i\Lambda_4(m_{123}\xi_{123}m_4\xi_4\xi_2 + m_1\xi_1 m_{234}\xi_{234}\xi_3)\\ &+\frac{i}{2}\Lambda_6(m_{12345}\xi_{12345}m_6\xi_6 - m_1\xi_1 m_{23456}\xi_{23456})\\ =&\frac{i}{2}\Lambda_4(\sigma_4(\xi_1,\xi_2,\xi_3,\xi_4)) + \frac{i}{6}\Lambda_6(\sigma_6(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5,\xi_6)),\end{aligned}$$

where

$$\sigma_4(\xi_1,\xi_2,\xi_3,\xi_4) = m_1^2\xi_1^2\xi_3 + m_2^2\xi_2^2\xi_4 + m_3^2\xi_3^2\xi_1 + m_4^2\xi_4^2\xi_2 \tag{21}$$

and

$$\sigma_6(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5,\xi_6) = \sum_{j=1}^{6}(-1)^{j-1}m_j^2\xi_j^2. \tag{22}$$

Notice that the contribution of Λ_2 is zero because the factor $(\xi_1^2-\xi_2^2)$ is zero over the set of integration $\xi_1+\xi_2=0$.

Differentiating $\Lambda_4(M_4)$, we have

$$\begin{aligned}&\frac{d}{dt}\Lambda_4(M_4(\xi_1,\xi_2,\xi_3,\xi_4))\\ &= i\Lambda_4\left(M_4\sum_{j=1}^{4}(-1)^j\xi_j^2\right)\\ &\quad - i\Lambda_6(M_4(\xi_{123},\xi_4,\xi_5,\xi_6)\xi_2 + M_4(\xi_1,\xi_{234},\xi_5,\xi_6)\xi_3\\ &\quad + M_4(\xi_1,\xi_2,\xi_{345},\xi_6)\xi_4 + M_4(\xi_1,\xi_2,\xi_3,\xi_{456})\xi_5)\\ &\quad + \frac{i}{2}\Lambda_8(M_4(\xi_{12345},\xi_6,\xi_7,\xi_8) - M_4(\xi_1,\xi_{23456},\xi_7,\xi_8)\\ &\quad + M_4(\xi_1,\xi_2,\xi_{34567},\xi_8) - M_4(\xi_1,\xi_2,\xi_3,\xi_{45678}))\\ &= i\Lambda_4\left(M_4\sum_{j=1}^{4}(-1)^j\xi_j^2\right)\\ &\quad - \frac{i}{36}\sum_{\substack{\{a,c,e\}=\{1,3,5\}\\ \{b,d,f\}=\{2,4,6\}}}\Lambda_6(M_4(\xi_{abc},\xi_d,\xi_e,\xi_f)\xi_b + M_4(\xi_a,\xi_{bcd},\xi_e,\xi_f)\xi_c\\ &\quad + M_4(\xi_a,\xi_b,\xi_{cde},\xi_f)\xi_d + M_4(\xi_a,\xi_b,\xi_c,\xi_{def})\xi_e)\\ &\quad + C\sum_{\substack{\{a,c,e,g\}=\{1,3,5,7\}\\ \{b,d,f,h\}=\{2,4,6,8\}}}\Lambda_8(M_4(\xi_{abcde},\xi_f,\xi_g,\xi_h) + M_4(\xi_a,\xi_b,\xi_{cdefg},\xi_h)\\ &\quad - M_4(\xi_a,\xi_{bcdef},\xi_g,\xi_h) - M_4(\xi_a,\xi_b,\xi_c,\xi_{defgh})).\end{aligned}$$

Then

$$\begin{aligned}\frac{d}{dt}E^2(w) = &-\frac{i}{2}\Lambda_4(\sigma_4(\xi_1,\xi_2,\xi_3,\xi_4)) + \frac{i}{2}\Lambda_4\left(M_4\sum_{j=1}^{4}(-1)^j\xi_j^2\right)\\ &-\frac{i}{6}\Lambda_6(\sigma_6(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5,\xi_6))\\ &-\frac{i}{72}\sum_{\substack{\{a,c,e\}=\{1,3,5\}\\ \{b,d,f\}=\{2,4,6\}}}\Lambda_6(M_4(\xi_{abc},\xi_d,\xi_e,\xi_f)\xi_b + M_4(\xi_a,\xi_{bcd},\xi_e,\xi_f)\xi_c\\ &+M_4(\xi_a,\xi_b,\xi_{cde},\xi_f)\xi_d + M_4(\xi_a,\xi_b,\xi_c,\xi_{def})\xi_e)\\ &+C_1\sum_{\substack{\{a,c,e,g\}=\{1,3,5,7\}\\ \{b,d,f,h\}=\{2,4,6,8\}}}\Lambda_8(M_4(\xi_{abcde},\xi_f,\xi_g,\xi_h) + M_4(\xi_a,\xi_b,\xi_{cdefg},\xi_h)\\ &-M_4(\xi_a,\xi_{bcdef},\xi_g,\xi_h) - M_4(\xi_a,\xi_b,\xi_c,\xi_{defgh})).\end{aligned}$$

We abbreviate the 6-linear and the 8-linear expressions as $\Lambda_6(M_6(\xi_1,\xi_2,\dots,\xi_6))$ and $\Lambda_8(M_8(\xi_1,\xi_2,\dots,\xi_8))$. We are now ready to make our choice for M_4. From our calculations in [9], we realized that the estimates for the different pieces of Λ_n appearing in the right-hand side of $\frac{d}{dt}E_N(w)$ are easier for n larger.[3] We decided to use the freedom of choosing M_4 to cancel the Λ_4 contribution obtained above. Hence, using (21), we set

$$M_4(\xi_1,\xi_2,\xi_3,\xi_4) = -\frac{m_1^2\xi_1^2\xi_3 + m_2^2\xi_2^2\xi_4 + m_3^2\xi_3^2\xi_1 + m_4^2\xi_4^2\xi_2}{\xi_1^2-\xi_2^2+\xi_3^2-\xi_4^2}, \tag{23}$$

which in the set of integration $\xi_1+\xi_2+\xi_3+\xi_4=0$ can also be written as

$$M_4(\xi_1,\xi_2,\xi_3,\xi_4) = -\frac{m_1^2\xi_1^2\xi_3 + m_2^2\xi_2^2\xi_4 + m_3^2\xi_3^2\xi_1 + m_4^2\xi_4^2\xi_2}{2\xi_{12}\xi_{14}}.$$

Remark 3.6. If we assume that $m(\xi)=1$, then $E^2(w)=E(w)$. In fact, on the set $\xi_1+\xi_2+\xi_3+\xi_4=0$ we have

$$\begin{aligned}&m_1^2\xi_1^2\xi_3 + m_2^2\xi_2^2\xi_4 + m_3^2\xi_3^2\xi_1 + m_4^2\xi_4^2\xi_2\\ &= \xi_1^2\xi_3 + \xi_2^2\xi_4 + \xi_3^2\xi_1 + \xi_4^2\xi_2\\ &= (\xi_1+\xi_3)(\xi_1\xi_3 - \xi_2\xi_4)\\ &= (\xi_1+\xi_3)(\xi_1\xi_3 + (\xi_1+\xi_3+\xi_4)\xi_4)\\ &= -(\xi_1+\xi_3)(\xi_1+\xi_4)(\xi_1+\xi_2);\end{aligned}$$

hence

$$M_4(\xi_1,\xi_2,\xi_3,\xi_4) = \frac{1}{2}(\xi_1+\xi_3) \tag{24}$$

and

$$E^2(w) = -\Lambda_2(\xi_1\xi_2) + \frac{1}{4}\Lambda_4(\xi_{13}),$$

[3] Compare, for example, sections 8, 9, and 10 in [9].

which is exactly the value of $E(w)$ in (15).

Once again we recall that we assume throughout the paper that $s \in (\frac{1}{2}, \frac{2}{3}]$ and that the multiplier m is defined as in (18). To stress the fact that with this choice the energy $E^2(w)$ depends on the parameter N, we write $E^2(w) = E^2_N$. We now summarize some of the above observations in the following.

PROPOSITION 3.7. *Let w be an H^1 global solution to* (12). *Then for any $T \in \mathbb{R}$ and $\delta > 0$ we have*

$$E^2_N(w(T+\delta)) - E^2_N(w(T)) = \int_T^{T+\delta} [\Lambda_6(M_6; w(t)) + \Lambda_8(M_8; w(t))]\, dt,$$

where the multipliers M_6 and M_8 are given by

$$\begin{aligned} M_6 :=& -\frac{i}{6}\sigma_6(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5,\xi_6) \\ &- \frac{i}{72} \sum_{\substack{\{a,c,e\}=\{1,3,5\}\\ \{b,d,f\}=\{2,4,6\}}} (M_4(\xi_{abc},\xi_d,\xi_e,\xi_f)\xi_b + M_4(\xi_a,\xi_{bcd},\xi_e,\xi_f)\xi_c \\ &+ M_4(\xi_a,\xi_b,\xi_{cde},\xi_f)\xi_d + M_4(\xi_a,\xi_b,\xi_c,\xi_{def})\xi_e), \\ M_8 :=& C_2 \sum_{\substack{\{a,c,e,g\}=\{1,3,5,7\}\\ \{b,d,f,h\}=\{2,4,6,8\}}} (M_4(\xi_{abcde},\xi_f,\xi_g,\xi_h) + M_4(\xi_a,\xi_b,\xi_{cdefg},\xi_h) \\ &- M_4(\xi_a,\xi_{bcdef},\xi_g,\xi_h) - M_4(\xi_a,\xi_b,\xi_c,\xi_{defgh})), \end{aligned}$$

where C_2 is an absolute constant. Furthermore, if $|\xi_j| \ll N$ for all j, then the multipliers M_6 and M_8 vanish.

We end this section with a lemma that shows the energy $E^2_N(w)$ has the same strength as $\|Iw\|_{H^1}$.

LEMMA 3.8. *Assume that w satisfies $\|w\|_{L^2} < \sqrt{2\pi}$, $\|Iw\|_{H^1} = O(1)$. Then, for $N \gg 1$,*

$$\|\partial_x Iw\|^2_{L^2} \lesssim E^2_N(w). \tag{25}$$

The proof of this lemma relies strongly on the estimate of the multiplier M_4, and it can be found in the next section.

4. Estimates for M_4 and proof of Lemma 3.8. Before we start with our estimates we recall some notation that we used in [9]. Let $n = 4$, 6, or 8 and let $\xi_1, \dots, \xi_n$ be frequencies such that $\xi_1 + \cdots + \xi_n = 0$. Define $N_i := |\xi_i|$, and $N_{ij} := |\xi_{ij}|$. We adopt the notation that

$$1 \le soprano, alto, tenor, baritone \le n$$

are the distinct indices such that

$$N_{soprano} \ge N_{alto} \ge N_{tenor} \ge N_{baritone}$$

are the highest, second highest, third highest, and fourth highest values of the frequencies $N_1, \dots, N_n$, respectively. (If there is a tie in frequencies, we break the tie arbitrarily.) Since $\xi_1 + \cdots + \xi_n = 0$, we must have $N_{soprano} \sim N_{alto}$. Also, from Proposition 3.7 we see that M_n vanishes unless $N_{soprano} \gtrsim N$.

In this section whenever we write $\max|f(\theta)|$ for a function f we understand that the maximum is taken for $|\theta| \sim N_{soprano}$.

LEMMA 4.1. *Assume M_4 is the multiplier defined in* (23) *and $m(\xi)$ is as in* (18). *Then*

$$|M_4(\xi_1, \dots, \xi_4)| \lesssim m^2(N_{soprano})N_{soprano}. \tag{26}$$

Proof. We observe that to prove (26) it suffices to prove

$$|\sigma_4(\xi_1, \dots, \xi_4)| \lesssim |\xi_{12}||\xi_{12}|m^2(N_{soprano})N_{soprano}.$$

Without loss of generality we may assume that $N_{soprano} = N_1$. By symmetry we can assume that $|\xi_{12}| \le |\xi_{14}|$. We divide the analysis into two cases: Case (a) when $N_1 \lesssim |\xi_{14}|$ and Case (b) when $|\xi_{14}| \ll N_1$.

Case (a). We write

$$\begin{aligned} |\sigma_4(\xi_1, \dots, \xi_4)| &= |m_1^2\xi_1^2\xi_3 + m_2^2\xi_2^2(-\xi_{12} - \xi_3) + m_3^2\xi_3^2\xi_1 + m_{12+3}^2\xi_{12+3}^2\xi_2| \\ &= |\xi_3(m_1^2\xi_1^2 - m_{1-12}^2\xi_{1-12}^2) + \xi_1(m_3^2\xi_3 - m_{3+12}^2\xi_{3+12}^2) \\ &\quad - \xi_{12}(m_2^2\xi_2^2 - m_{12+3}^2\xi_{12+3}^2)|. \end{aligned} \tag{27}$$

Then the MVT shows that

$$|\sigma_4(\xi_1, \xi_2, \xi_3, \xi_4)| \lesssim |\xi_{12}|N_1 \max|(m(\xi)^2\xi^2)'|, \tag{28}$$

where $|\xi| \lesssim N_1$. Now it is easy to see that for m defined in (18)

$$(m^2(\xi)\xi^2)' \sim m^2(\xi)\xi$$

and that the function $m^2(\xi)\xi$ is nondecreasing. Then (28) immediately gives (26).

Case (b). We first write σ_4 so that the DMVT in Lemma 2.3 can be applied. For simplicity we write $m^2(\xi)\xi^2 = f(\xi)$. Then in the set $\xi_1 + \cdots + \xi_4 = 0$ we have

$$\begin{aligned} \sigma_4(\xi_1, \dots, \xi_4) &= f(\xi_1)\xi_3 + f(\xi_2)\xi_4 + f(\xi_3)\xi_1 f(\xi_4)\xi_2 \\ &= \xi_3[f(\xi_1) - f(\xi_2)] + \xi_1[f(\xi_3) - f(-\xi_4)] - \xi_{12}[f(\xi_2) - f(-\xi_4)] \\ &= \xi_3[f(\xi_1) - f(\xi_2) + f(\xi_3) - f(-\xi_4)] \\ &\quad + (\xi_1 - \xi_3)[f(\xi_3) - f(\xi_3 - \xi_{12})] - \xi_{12}[f(\xi_2) - f(-\xi_4)] \\ &= \xi_3[f(\xi_1 - \xi_{12} - \xi_{14}) - f(\xi_1 - \xi_{12}) - f(\xi_1 - \xi_{14}) + f(\xi_1)] \\ &\quad + (-\xi_3 + \xi_1)[f(\xi_3) - f(\xi_3 - \xi_{12})] - \xi_{12}[f(\xi_2) - f(\xi_2 + \xi_{14} - \xi_{12})], \end{aligned}$$

where we often used the fact that $f(\xi)$ is an even function. Using the DMVT in the first term of the right-hand side of the inequality and the MVT in the remaining two terms we obtain

$$\sigma_4(\xi_1, \dots, \xi_4) \lesssim |\xi_1||f''(\theta)||\xi_{12}||\xi_{14}| + |\xi_{12}| \max|f'|(|\xi_{3-1}| + |\xi_{14}| + |\xi_{12}|), \tag{29}$$

where $|\theta| \sim N_1$. Now observe that

$$|\xi_{3-1}| = |\xi_{12} + \xi_{14}| \lesssim |\xi_{14}|$$

and that $|f''(\theta)| \lesssim m(N_1)^2$, so inserting (29) in the definition of M_4 we obtain (26). ∎

We need two more local estimates for M_4.

LEMMA 4.2.

- *Assume that* $|\xi_1| \sim |\xi_3| \gtrsim N \gg |\xi_2|, |\xi_4|$*; then*

$$|M_4(\xi_1, \xi_2, \xi_3, \xi_4)| \lesssim m(N_{soprano})^2 N_{tenor}. \tag{30}$$

- *Assume that* $|\xi_1| \sim |\xi_2| \gtrsim N \gg |\xi_3|, |\xi_4|$*; then*

$$M_4(\xi_1, \xi_2, \xi_3, \xi_4) = \frac{m_1^2 \xi_2^2}{2\xi_1} + R(\xi_1, \dots, \xi_4), \tag{31}$$

where

$$|R(\xi_1, \dots, \xi_4)| \lesssim N_{tenor}.$$

Proof. The first part of the lemma follows from the MVT. In fact,

$$\begin{aligned}\left| \frac{m_1^2 \xi_1^2 \xi_3 + \xi_2^2 \xi_4 + m_3^2 \xi_3^2 \xi_1 + \xi_4^2 \xi_2}{\xi_{12} \xi_{14}} \right| &\lesssim \frac{|\xi_1 \xi_3 \xi_{13}| \max |(m(\xi)^2 \xi)'| + |\xi_{24} \xi_2 \xi_4|}{|\xi_1|^2} \\ &\lesssim m(N_{soprano})^2 N_{tenor},\end{aligned}$$

where again we used that $|(m(\xi)^2 \xi)'| \sim |m(\xi)\xi|$.

To prove the second part of the lemma we use the identity

$$\frac{1}{\xi_{14}} = \frac{1}{\xi_1} - \frac{\xi_4}{\xi_{14}} \frac{1}{\xi_1},$$

and we write

$$-2M_4(\xi_1, \xi_2, \xi_3, \xi_4) + \frac{m_1^2 \xi_2^2}{\xi_1} = R_1(\xi_1, \dots, \xi_4) + R_2(\xi_1, \dots, \xi_4),$$

where

$$\begin{aligned}R_1(\xi_1, \dots, \xi_4) &= \frac{m_1^2 \xi_1^2 \xi_3 + m_2^2 \xi_2^2 \xi_4 + \xi_3^2 \xi_1 + \xi_4^2 \xi_2 + m_1^2 \xi_2^2 \xi_{12}}{\xi_{12} \xi_1}, \\ R_2(\xi_1, \dots, \xi_4) &= -\frac{\xi_4}{\xi_{14}} \frac{m_1^2 \xi_1^2 \xi_3 + m_2^2 \xi_2^2 \xi_4 + \xi_3^2 \xi_1 + \xi_4^2 \xi_2}{\xi_{12} \xi_1}.\end{aligned}$$

We first estimate R_1:

$$\begin{aligned}R_1(\xi_1, \dots, \xi_4) &= \frac{m_1^2 \xi_1^2 \xi_3 + m_2^2 \xi_2^2 \xi_4 + \xi_3^2 \xi_1 + \xi_4^2 \xi_2 - m_1^2 \xi_2^2 \xi_{34}}{\xi_{12} \xi_1} \\ &= \frac{m_1^2 \xi_3 (\xi_1^2 - \xi_2^2) + \xi_2^2 \xi_4 (m_2^2 - m_1^2) + \xi_3^2 (\xi_1 + \xi_2) + \xi_2 (\xi_4^2 - \xi_3^2)}{\xi_{12} \xi_1}, \text{ and}\end{aligned}$$

hence, by the MVT,

$$|R_1(\xi_1, \dots, \xi_4)| \lesssim N_{tenor}.$$

On the other hand,

$$R_2(\xi_1, \dots, \xi_4) = -\frac{\xi_4}{\xi_{14}} \frac{m_1^2 \xi_1^2 (\xi_3 + \xi_4) + (m_2^2 \xi_2^2 - m_1^2 \xi_1^2) \xi_4 + \xi_3^2 \xi_{12} + \xi_2 \xi_{34} \xi_{3-4}}{\xi_{12} \xi_1}, \text{ and}$$

hence, again by the MVT,

$$|R_2(\xi_1, \dots, \xi_4)| \lesssim N_{tenor}. \qquad \square$$

Proof of Lemma 3.8.
Proof. We rewrite $E_N^2(w)$ as

$$E_N^2(w) = -\Lambda_2(m_1\xi_1 m_2\xi_2) + \frac{1}{8}\Lambda_4(\xi_{13-24}m_1m_2m_3m_4) + \frac{1}{8}\Lambda_4(4M_4(\xi_1,\xi_2,\xi_3,\xi_4) - \xi_{13-24}m_1m_2m_3m_4).$$

In Lemma 3.6 of [9] we proved the estimate

$$\|\partial_x Iw\|_{L^2}^2 \lesssim -\Lambda_2(m_1\xi_1 m_2\xi_2) + \frac{1}{8}\Lambda_4(\xi_{13-24}m_1m_2m_3m_4)$$

for $\|Iw\|_{L^2} < \sqrt{2\pi}$. Hence we have only to show that

$$|\Lambda_4(4M_4(\xi_1,\xi_2,\xi_3,\xi_4) - \xi_{13-24}m_1m_2m_3m_4)| \lesssim O\left(\frac{1}{N^\alpha}\right)\|Iw\|_{H^1}^4 \tag{32}$$

for some $\alpha > 0$.

We first perform a Littlewood–Paley decomposition of the four factors w so that the ξ_i are essentially the constants N_i, $i = 1,\dots,4$. To recover the sum at the end we borrow a $N_{soprano}^{-\epsilon}$ from the large denominator $N_{soprano}$ and often this will not be mentioned.

If all $|\xi_j|$ are less than $\frac{N}{100}$, the left-hand side of (32) vanishes thanks to (23). Therefore, we may assume $N_{soprano} \gtrsim N$. Also note $N_{alto} \gtrsim N$ on the set $\xi_1+\xi_2+\xi_3+\xi_4 = 0$. Then it is obvious that

$$|\Lambda_4(\xi_{13-24}m_1m_2m_3m_4)| \lesssim \frac{1}{N}\|Iw\|_{H^1}^2\|Iw\|_{L^\infty}^2 \lesssim \frac{1}{N}\|Iw\|_{H^1}^4.$$

Next we control the contribution of $\Lambda_4(M_4)$ in (32). By (26), we have

$$|\Lambda_4(M_4(\xi_1,\xi_2,\xi_3,\xi_4))| \lesssim \frac{1}{N_{soprano}^{1-}m(N_{baritone})^2N_{baritone}}\|Iw\|_{H^1}^4 \lesssim \frac{1}{N^{1-}}\|Iw\|_{H^1}^4,$$

where again we used the fact that $m^2(\xi)\xi$ is nondecreasing. □

5. Local estimates. This section contains a refinement of the results presented in section 5 of [9]. We start with the main result.

THEOREM 5.1. *Let w be a H^1 global solution to* (12) *and let $T \in \mathbb{R}$ be such that*

$$\|Iw(T)\|_{H^1} \le C_0$$

for some $C_0 > 0$. Then we have

$$\|Iw\|_{X^{1,b}([T,T+\delta]\times\mathbb{R})} \lesssim 1$$

for any $\frac{1}{2} < b < \frac{3}{4}$ and for some $\delta > 0$ depending on C_0.

Remark 5.2. This theorem is stronger than the corresponding Theorem 5.1 in [9] because b can be arbitrarily close to $\frac{3}{4}$, and this is essential to obtain our sharp global well-posedness result.

As explained in [9] the proof of Theorem 5.1 is a consequence of the following multilinear estimates.

LEMMA 5.3. *For the Schwartz function w and $\frac{1}{2} < b < \frac{3}{4},\ b' < \frac{3}{4}$, we have*

$$\|I(w\partial_x \overline{w} w)\|_{1,b'-1} \lesssim \|Iw\|^2_{1,\frac{1}{2}+}\|Iw\|_{1,b}, \tag{33}$$

$$\|I(w\overline{w}w\overline{w}w)\|_{1,b'-1} \lesssim \|Iw\|^5_{1,\frac{1}{2}+}. \tag{34}$$

Proof. The proof of (34) follows from the same arguments used to prove (17) in [9], and we do not present it here again. The proof of (33) on the other hand is more delicate than the one given in [9] for (16), so we decided to give all the details. By standard duality arguments in L^2 and renormalization, it is easy to see that (33) is equivalent to

$$\int_* \frac{m_4\langle\xi_4\rangle|\xi_2|\langle\tau_4+\xi_4^2\rangle^{b'-1}}{\sum_{i=1}^3\langle\tau_i+(-1)^i\xi_i^2\rangle^{b-\frac{1}{2}-}\prod_{j=1}^3 m_j\langle\xi_j\rangle\langle\tau_j+(-1)^j\xi_j^2\rangle^{\frac{1}{2}+}}\prod_{j=1}^4 F_j(\tau_j,\xi_j) \lesssim \prod_{j=1}^4\|F_j\|_{L^2}, \tag{35}$$

where all functions F_j are real-valued and nonnegative. If

$$\frac{m_4\langle\xi_4\rangle|\xi_2|}{\prod_{j=1}^3 m_j\langle\xi_j\rangle} \lesssim 1, \tag{36}$$

then the L^2 estimate (5) for F_4 and the Strichartz estimate (6) with $p=6$ for F_1, F_2, F_3 automatically shows (35) for $b > \frac{1}{2},\ b' \le 1$. Then we may assume

$$\frac{m_4\langle\xi_4\rangle|\xi_2|}{\prod_{j=1}^3 m_j\langle\xi_j\rangle} \gg 1,$$

which, one can easily check, can happen only when

$$|\xi_2| \gg 1, \quad |\xi_{12}| \gg 1, \quad |\xi_{14}| \gg 1.$$

We recall (cf. [3] and [9]) the fundamental inequality

$$|\xi_{12}\xi_{14}| \lesssim \max_{j=1,2,3,4}\{\langle\tau_j+(-1)^j\xi_j^2\rangle\}. \tag{37}$$

Then we proceed with a case by case analysis: Case (a) if $\max_{j=1,2,3}\{\langle\tau_4+\xi_4^2\rangle, \langle\tau_j+(-1)^j\xi_j^2\rangle\} = \langle\tau_4+\xi_4^2\rangle$ and Case (b) if $\max_{j=1,2,3}\{\langle\tau_4+\xi_4^2\rangle, \langle\tau_j+(-1)^j\xi_j^2\rangle\} = \langle\tau_i+(-1)^j\xi_i^2\rangle$ for some $i=1,2,3$.

- *Case* (a). In this case we replace in the denominator $\langle\tau_4+\xi_4^2\rangle^{1-b'}$ with $(\langle\xi_{12}\rangle\langle\xi_{14}\rangle)^{1-b'}$. Then, using the same argument that in [9] led us from (16) to (18), we can show that (35) is equivalent to

$$\int_* \frac{\langle\xi_4\rangle^s\langle\xi_2\rangle^{1-s}}{(\langle\xi_{12}\rangle\langle\xi_{14}\rangle)^{1-b'}\langle\xi_1\rangle^s\langle\xi_3\rangle^s\prod_{j=1}^3\langle\tau_j+(-1)^j\xi_j^2\rangle^{\frac{1}{2}+}}\prod_{j=1}^4 F_j(\tau_j,\xi_j) \lesssim \prod_{j=1}^4\|F_j\|_{L^2}. \tag{38}$$

To have an idea of the "numerics" involved while proceeding with the proof, the reader should keep in mind that the interesting case is when $s = \frac{1}{2}+$ and $1 - b' = \frac{1}{4}+$. Since $\xi_{14} = -\xi_{32}$, by symmetry, we may assume that $|\xi_1| \geq |\xi_3|$. Then, using the fact that $\xi_4 = -\xi_3 - \xi_{12}$, we can write

$$\frac{\langle \xi_4 \rangle^s \langle \xi_2 \rangle^{1-s}}{(\langle \xi_{12} \rangle \langle \xi_{14} \rangle)^{1-b'} \langle \xi_1 \rangle^s \langle \xi_3 \rangle^s} = A_1 + A_2, \tag{39}$$

where

$$A_1 \lesssim \frac{\langle \xi_2 \rangle^{1-s}}{(\langle \xi_{12} \rangle \langle \xi_{14} \rangle)^{1-b'} \langle \xi_1 \rangle^s},$$
$$A_2 \lesssim \frac{\langle \xi_{12} \rangle^{s-1+b'} \langle \xi_2 \rangle^{1-s}}{\langle \xi_{14} \rangle^{1-b'} \langle \xi_1 \rangle^s \langle \xi_3 \rangle^s}.$$

We now write $\xi_{12} = -\xi_{14} - \xi_3 + \xi_1$, and we write

$$A_2 = A_2^1 + A_2^2 + A_2^3,$$

where

$$A_2^1 \lesssim \frac{\langle \xi_2 \rangle^{1-s}}{\langle \xi_{14} \rangle^{2(1-b')-s} \langle \xi_1 \rangle^s \langle \xi_3 \rangle^s},$$
$$A_2^2 \lesssim \frac{\langle \xi_2 \rangle^{1-s}}{\langle \xi_{14} \rangle^{1-b'} \langle \xi_3 \rangle^{1-b'} \langle \xi_1 \rangle^s},$$
$$A_2^3 \lesssim \frac{\langle \xi_2 \rangle^{1-s}}{\langle \xi_{14} \rangle^{1-b'} \langle \xi_1 \rangle^{1-b'} \langle \xi_3 \rangle^s}.$$

It is now easy to see that, for $1 - b' \geq \frac{s}{2}$,

$$A_1, A_2^i(\xi_1, \xi_2, \xi_3) \lesssim \frac{\langle \xi_2 \rangle^{\frac{1}{2}}}{\langle \xi_1 \rangle^{\frac{s}{2}} \langle \xi_3 \rangle^{\frac{s}{2}}} \quad \text{for all } i = 1, 2, 3.$$

Then by (9) and (11) we obtain

$$\int_* \frac{\langle \xi_4 \rangle^s \langle \xi_2 \rangle^{1-s}}{(\langle \xi_{12} \rangle \langle \xi_{14} \rangle)^{1-b'} \langle \xi_1 \rangle^s \langle \xi_3 \rangle^s \prod_{j=1}^3 \langle \tau_j + (-1)^j \xi_j^2 \rangle^{\frac{1}{2}+}} \prod_{j=1}^4 F_j(\tau_j, \xi_j)$$
$$\lesssim \|\widetilde{\mathcal{F}}^{-1}(F_4)\|_{L^2_{xt}} \left\| \widetilde{\mathcal{F}}^{-1} \left(\frac{\langle \xi \rangle^{\frac{1}{2}}}{\langle \tau + \xi^2 \rangle^{\frac{1}{2}+}} F_2 \right) \right\|_{L^\infty_x L^2_t} \| \widetilde{\mathcal{F}}^{-1} \left(\frac{\langle \xi \rangle^{-\frac{s}{2}}}{\langle \tau - \xi^2 \rangle^{\frac{1}{2}+}} F_3 \right) \|_{L^4_x L^\infty_t}$$
$$\times \| \widetilde{\mathcal{F}}^{-1} \left(\frac{\langle \xi \rangle^{-\frac{s}{2}}}{\langle \tau - \xi^2 \rangle^{\frac{1}{2}+}} F_1 \right) \|_{L^4_x L^\infty_t} \lesssim \prod_{j=1}^4 \|F_j\|_{L^2}.$$

- *Case* (b). In this case we borrow a power $\alpha = b' - \frac{1}{2}+$ from the large denominator, and we reduce our estimate to

$$\int_* \frac{\langle \xi_4 \rangle^s \langle \xi_2 \rangle^{1-s}}{\langle \xi_1 \rangle^s \langle \xi_3 \rangle^s \prod_{j=1}^4 \langle \tau_j + (-1)^j \xi_j^2 \rangle^{\frac{1}{2}+}} \prod_{j=1}^4 F_j(\tau_j, \xi_j) \lesssim \prod_{j=1}^4 \|F_j\|_{L^2}.$$

Again by symmetry we can assume that $|\xi_1| \geq |\xi_3|$. We first observe that if the exponent of $\langle \xi_4 \rangle$ were $\frac{1}{2}$, then we could simply use (9) for the function F_2 and (10) for the function F_4 to obtain the estimate as we did above. However, in our case $s > \frac{1}{2}$, so we have to do a bit more work. We subdivide the analysis into subcases.

– *Subcase* (1). $|\xi_4| \lesssim |\xi_2|$. In this case we can write

$$\langle \xi_4 \rangle^s \langle \xi_2 \rangle^{1-s} \lesssim \langle \xi_4 \rangle^{\frac{1}{2}} \langle \xi_2 \rangle^{\frac{1}{2}},$$

and we can indeed use (9) and (10).

– *Subcase* (2). $|\xi_2| \ll |\xi_4|$. Because we assumed that $|\xi_3| \leq |\xi_1|$ and we are on the set $\xi_1 + \cdots + \xi_4 = 0$, it follows that $|\xi_4| \lesssim |\xi_1|$. Then the estimate becomes

$$\begin{aligned}
&\int_* \frac{\langle \xi_2 \rangle^{1-s}}{\langle \xi_3 \rangle^s \prod_{j=1}^4 \langle \tau_j + (-1)^j \xi_j^2 \rangle^{\frac{1}{2}+}} \prod_{j=1}^4 F_j(\tau_j, \xi_j) \\
&\lesssim \left\| \widetilde{\mathcal{F}}^{-1} \left(\frac{1}{\langle \tau + \xi^2 \rangle^{\frac{1}{2}+}} F_4 \right) \right\|_{L^4_{xt}} \left\| \widetilde{\mathcal{F}}^{-1} \left(\frac{1}{\langle \tau - \xi^2 \rangle^{\frac{1}{2}+}} F_1 \right) \right\|_{L^4_{xt}} \\
&\times \left\| \widetilde{\mathcal{F}}^{-1} \left(\frac{\langle \xi \rangle^{1-s}}{\langle \tau + \xi^2 \rangle^{\frac{1}{2}+}} F_2 \right) \right\|_{L^\infty_x L^2_t} \\
&\times \left\| \widetilde{\mathcal{F}}^{-1} \left(\frac{\langle \xi \rangle^{-s}}{\langle \tau - \xi^2 \rangle^{\frac{1}{2}+}} F_3 \right) \right\|_{L^2_x L^\infty_t} \lesssim \prod_{j=1}^4 \|F_j\|_{L^2},
\end{aligned}$$

thanks to (6) for $p = 2$, (9), and (10). □

6. Proof of Proposition 3.2. Based on Lemma 3.8, Theorem 5.1, and the arguments presented in [9, section 6] (see also the comments in [9, section 7]), the only result that one needs to obtain is the following.

LEMMA 6.1. *For any Schwartz function w, we have*

$$\left| \int_T^{T+\delta} \Lambda_n(M_n; w(t))\, dt \right| \lesssim \frac{1}{N^{2-}} \|Iw\|^n_{X^{1,3/4-}([T,T+\delta]\times\mathbb{R})} \tag{40}$$

for $n = 6, 8$, where M_6, M_8 are defined in Proposition 3.7.

In [9] we were only able to obtain a decay of N^{-1+}, which is why we could only prove global well-posedness for $s > \frac{2}{3}$.

The proof of this lemma is a corollary of the four lemmas that follow in this section.

LEMMA 6.2 ($n = 8$).

$$|M_8(\xi_1, \xi_2, \ldots, \xi_8)| \lesssim N_{soprano} m^2(N_{soprano}).$$

This is a simple consequence of Lemma 4.1. We now turn to the estimate of $\frac{d}{dt} E^2(Iw)$ involving Λ_8.

LEMMA 6.3.

$$\left| \int_T^{T+\delta} \int \Lambda_8(M_8(\xi_1, \xi_2, \ldots, \xi_8))\, dt \right| \lesssim \frac{1}{N^{2-}} \|Iw\|^8_{1,\frac{1}{2}+}.$$

Proof. As in the proof of Lemma 3.8, also in this case we first perform a Littlewood–Paley decomposition of the eight factors w so that the ξ_i essentially are the constants N_i, $i = 1, \dots, 8$. To recover the sum at the end we borrow a $N_{soprano}^{-\epsilon}$ from the large denominator $N_{soprano}$. Often this will not be mentioned, and it will only be recorded at the end by paying a price equivalent to N^{0+}. Below we often use the set of indices $R = \{soprano, alto, tenor\}$. Again we proceed by analyzing different cases.

- *Case* (a). $N_{soprano} \sim N_{tenor}$. By Lemma 6.2 and the fact that $m(\xi)\langle\xi\rangle^{\frac{1}{2}}$ is increasing, we have

$$\begin{aligned}&\left|\int_T^{T+\delta}\int \Lambda_8(M_8(\xi_1,\xi_2,\dots,\xi_8))\,dt\right|\\ &\lesssim \sum_{R,j}\frac{N_{soprano}}{m(N_{tenor})}\|D_xIw_{soprano}\|_{L^6}\|D_xIw_{alto}\|_{L^6}\|D_xIw_{tenor}\|_{L^6}\\ &\quad\prod_{j,k\notin R}\|D_xIw_j\|_{L^6}\|D_x^{1/2-}Iw_k\|_{L^\infty}^2 \lesssim \frac{1}{N^{2-}}\|Iw\|_{1,\frac{1}{2}+}^8.\end{aligned}$$

- *Case* (b). $N_{soprano} \gg N_{tenor}$. By Lemma 2.2, and again the monotonicity of $m(\xi)\langle\xi\rangle^{1/2}$, we have

$$\begin{aligned}&\left|\int_T^{T+\delta}\int \Lambda_8(M_8(\xi_1,\xi_2,\dots,\xi_8))\,dt\right|\\ &\lesssim N_{soprano}\|Iw_{soprano}w_{tenor}\|_{L^2}\|Iw_{alto}w_{baritone}\|_{L^2}\\ &\times\|w\|_{L^\infty}^4 \lesssim \frac{1}{N^{2-}}\|Iw\|_{1,\frac{1}{2}+}^8. \qquad \square\end{aligned}$$

LEMMA 6.4 ($n = 6$).

- *If* $N_{tenor} \gtrsim N$, *we have*

$$|M_6(\xi_1,\xi_2,\dots,\xi_6)| \lesssim m(N_{soprano})^2 N_{soprano}^2. \tag{41}$$

- *If* $N_{tenor} \ll N$, *we have*

$$|M_6(\xi_1,\xi_2,\dots,\xi_6)| \lesssim N_{soprano}N_{tenor}. \tag{42}$$

Proof. If $N_{soprano} \ll N$, M_6 vanishes. Then we may assume $N_{soprano} \gtrsim N$. Also in the set $\xi_1 + \cdots + \xi_6 = 0$ we have $N_{alto} \sim N_{soprano}$.

The proof of (41) follows from (26). The proof of (42) is more delicate. By symmetry we assume $soprano = 1$, $N_1 \geq N_3 \geq N_5$, $N_2 \geq N_4 \geq N_6$. Again we analyze different cases.

- *Case* (a). $alto = 2$. The MVT shows

$$\begin{aligned}|\sigma_6(\xi_1,\xi_2,\dots,\xi_6)| &\lesssim m(N_1)^2N_1N_{12} + m(N_{tenor})^2N_{tenor}^2\\ &\lesssim m(N_{soprano})^2N_{soprano}N_{tenor}.\end{aligned}$$

Next we estimate the second term in M_6:

$$\sum (M_4(\xi_{abc},\xi_d,\xi_e,\xi_f)\xi_b + M_4(\xi_a,\xi_{bcd},\xi_e,\xi_f)\xi_c + M_4(\xi_a,\xi_b,\xi_{cde},\xi_f)\xi_d + M_4(\xi_a,\xi_b,\xi_c,\xi_{def})\xi_e).$$

Again by (26) one has that

$$|M_4(\xi_{abc},\xi_d,\xi_e,\xi_f)\xi_g| \lesssim m(N_{soprano})^2 N_{soprano} N_{tenor} \tag{43}$$

for every $a,\dots,g \in \{1,\dots,6\}$ and $g \neq soprano, alto$. Thus we have only to consider the contributions

$$\left| \sum_{(a,e)\in\{3,5\}} \sum_{(d,f)\in\{4,6\}} M_4(\xi_{a21},\xi_d,\xi_e,\xi_f)\xi_2 + M_4(\xi_a,\xi_{21d},\xi_e,\xi_f)\xi_1 \right|$$

$$+ \left| \sum_{(a,c)\in\{3,5\}} \sum_{(d,f)\in\{4,6\}} M_4(\xi_a,\xi_{12b},\xi_e,\xi_f)\xi_1 + M_4(\xi_a,\xi_b,\xi_{12e},\xi_f)\xi_2 \right|$$

$$+ \left| \sum_{(a,c)\in\{3,5\}} \sum_{(d,f)\in\{4,6\}} M_4(\xi_a,\xi_b,\xi_{12c},\xi_f)\xi_2 + M_4(\xi_a,\xi_b,\xi_c,\xi_{12f})\xi_1 \right|$$

$$+ \left| \sum_{(a,e)\in\{3,5\}} \sum_{(d,f)\in\{4,6\}} M_4(\xi_{a2c},\xi_d,\xi_1,\xi_f)\xi_2 + M_4(\xi_a,\xi_2,\xi_c,\xi_{d1f})\xi_1 \right| = \sum_{i=1}^{4} I_i.$$

Observe first that all the variables appearing in the function M_4 in $\sum_{i=1}^3 I_i$ are strictly smaller that $\frac{N}{2}$, and hence by (24) it follows that

$$\sum_{i=1}^{3} I_i \lesssim N_{soprano} N_{tenor}.$$

To estimate I_4 we use (30) and the symmetry of M_4. Then also in this case we obtain

$$I_4 \lesssim N_{soprano} N_{tenor}.$$

- *Case* (b). $alto = 3$. In this case we need some cancellation between the large terms coming from $\sigma_6(\xi_1,\dots,\xi_6)$ and the large terms of the sum of the M_4. From (43) it is easy to see that one needs to estimate only

$$\begin{aligned}\widetilde{M}_6(\xi_1,\dots,\xi_6) = &-\frac{1}{6}(m_1^2\xi_1^2 + m_3^2\xi_3^2) \\ &- \frac{\xi_1}{36}\left(\sum_{(b,d,f)\in\{2,4,6\}} M_4(\xi_a,\xi_{b1d},\xi_3,\xi_f) + M_4(\xi_a,\xi_b,\xi_3,\xi_{d1f})\right) \\ &- \frac{\xi_3}{36}\left(\sum_{(b,d,f)\in\{2,4,6\}} M_4(\xi_a,\xi_b,\xi_1,\xi_{d3f}) + M_4(\xi_a,\xi_{b3d},\xi_1,\xi_f)\right).\end{aligned}$$

We now use (31) and the symmetries of M_4 to write

$$\begin{aligned}
\widetilde{M_6}(\xi_1,\dots,\xi_6) = &-\frac{1}{6}(m_1^2\xi_1^2+m_3^2\xi_3^2)\\
&-\frac{\xi_1}{72}\left(\sum_{(b,d,f)\in\{2,4,6\}}\frac{m_3^2(\xi_{b1d}^2+\xi_{b1f}^2)}{\xi_3}\right)+O(N_{soprano}N_{tenor})\\
&-\frac{\xi_3}{72}\left(\sum_{(b,d,f)\in\{2,4,6\}}\frac{m_1^2(\xi_{d3f}^2+\xi_{b3d}^2)}{\xi_1}\right)+O(N_{soprano}N_{tenor})\\
=&-\frac{1}{6}(m_1^2\xi_1^2+m_3^2\xi_3^2)\\
&+\frac{1}{72}\left(\sum_{(b,d,f)\in\{2,4,6\}}m_3^2(\xi_{b1d}^2+\xi_{b1f}^2)\right)+O(N_{soprano}N_{tenor})\\
&+\frac{1}{72}\left(\sum_{(b,d,f)\in\{2,4,6\}}m_1^2(\xi_{d3f}^2+\xi_{b3d}^2)\right)+O(N_{soprano}N_{tenor})\\
=&-\frac{1}{72}m_3^2\sum_{(b,d,f)\in\{2,4,6\}}(\xi_3^2-\xi_{1bd}^2)+(\xi_3^2-\xi_{1fb}^2)\\
&-\frac{1}{72}m_1^2\sum_{(b,d,f)\in\{2,4,6\}}(\xi_1^2-\xi_{3bf}^2)+(\xi_1^2-\xi_{b3d}^2)\\
&+O(N_{soprano}N_{tenor}),
\end{aligned}$$

and now it is clear that also in this case

$$|\widetilde{M_6}(\xi_1,\dots,\xi_6)|\lesssim N_{soprano}N_{tenor}. \qquad \square$$

LEMMA 6.5.

$$\left|\int_T^{T+\delta}\int\Lambda_6(M_6(\xi_1,\xi_2,\dots,\xi_6))\,dt\right|\lesssim\frac{1}{N^{2-}}\|Iw\|_{1,\frac{3}{4}-}^6. \tag{44}$$

Proof. Also in this case one uses a Littlewood–Paley decomposition to start. We divide the proof into three different cases: Case (a) when $N_{baritone}\gtrsim N$, Case (b) when $N_{soprano}\geq N_{tenor}\gtrsim N\gg N_{baritone}$, and Case (c) when $N_{soprano}\sim N_{alto}\gtrsim N\gg N_{tenor}$. Below we often use the two sets of indices $S=\{soprano, alto, tenor, baritone\}$ and $R=\{soprano, alto, tenor\}$. We also recall that thanks to the fact that $m(\xi)|\xi|^{\frac{1}{2}}$ is not decreasing,

$$m(\xi)(1+|\xi|)\gtrsim\begin{cases}N & \text{if } |\xi|>\frac{N}{2},\\ 1 & \text{if } |\xi|\leq\frac{N}{2}.\end{cases} \tag{45}$$

- *Case* (a). $N_{baritone} \gtrsim N$. By Lemma 6.4, (45), and the Strichartz estimate (4), we have

$$\begin{aligned}
&\left| \int_T^{T+\delta} \int \Lambda_6(M_6(\xi_1, \xi_2, \dots, \xi_6))\, dt \right| \\
&\lesssim \sum_{S,j} \frac{1}{N_{soprano} N} m(N_{soprano}) N_{soprano} \|w_{soprano}\|_{L^6} \\
&\times m(N_{alto}) N_{alto} \|w_{alto}\|_{L^6} m(N_{tenor}) N_{tenor} \|w_{tenor}\|_{L^6} \\
&\times m(N_{baritone}) N_{baritone} \|w_{baritone}\|_{L^6} \\
&\times \prod_{j \notin S} \|Iw_j\|_{L^6} \lesssim \frac{1}{N^{2-}} \|Iw\|^6_{1,\frac{1}{2}+}.
\end{aligned}$$

- *Case* (b). $N_{soprano} \geq N_{tenor} \gtrsim N \gg N_{baritone}$. This is the only part in which we need to use the space $X^{1,b}$ with $b \sim \frac{3}{4}-$. By Lemma 6.4 and (45) we have

$$\begin{aligned}
&\left| \int_T^{T+\delta} \int \Lambda_6(M_6(\xi_1, \xi_2, \dots, \xi_6))\, dt \right| \\
&\lesssim \sum_{R,j} \frac{1}{N_{soprano}} m(N_{soprano}) N_{soprano} \|w_{soprano} w_{baritone}\|_{L^2} \\
&\times m(N_{alto}) N_{alto} \|w_{alto}\|_{L^6} m(N_{tenor}) N_{tenor} \|w_{tenor}\|_{L^6} \prod_{j \notin R} \|D_x^{\frac{1}{2}} Iw_j\|_{L^{12}}.
\end{aligned}$$

Using Lemma 2.2 and (45), it is easy to see that

$$\begin{aligned}
&m(N_{soprano}) N_{soprano} \|w_{soprano} w_{baritone}\|_{L^2} \\
&\lesssim N^{-1/2} \|Iw_{soprano}\|_{X^{1,\frac{1}{2}+}} \|Iw_{baritone}\|_{X^{1,\frac{1}{2}+}}.
\end{aligned}$$

Also by the Sobolev inequalities and again (45)

$$\prod_{j \notin R} \|D_x^{\frac{1}{2}} Iw_j\|_{L^{12}} \lesssim \prod_{j \notin R} \|Iw_j\|_{X^{1,\frac{1}{2}+}}.$$

Collecting the above estimates one obtains

$$\left| \int_T^{T+\delta} \int \Lambda_6(M_6(\xi_1, \xi_2, \dots, \xi_6))\, dt \right| \lesssim \frac{1}{N^{\frac{3}{2}-}} \|Iw\|^6_{1,\frac{1}{2}+}.$$

Unfortunately, the decay $N^{-\frac{3}{2}+}$ is not enough for our purposes. Because the local estimate allow us to handle terms of type $\|Iw\|_{1,\frac{3}{4}-}$ (see section 5), we take advantage of the extra denominators. To see this we use the identity

$$\xi_1 + \cdots + \xi_4 = 0 \implies \xi_1^2 - \xi_2^2 + \xi_3^2 - \xi_4^2 = 2\xi_{12}\xi_{14},$$

proved in [9]. We consider only the case $N_1 = N_{soprano}, N_2 = N_{alto}$, and $N_3 = N_{tenor}$. Indeed if $N_5 = N_{tenor}$ the argument is easier. Then in the set $\xi_1 + \cdots + \xi_6 = 0$ we write

$$\begin{aligned}\sum_{i=1}^{6}(-1)^{i-1}\xi_i^2 &= \xi_1^2 - \xi_2^2 + \xi_3^2 - (\xi_4+\xi_5+\xi_6)^2 \\ &\quad +(\xi_4+\xi_5+\xi_6)^2 - \xi_4^2 + \xi_5^2 - \xi_6^2 \\ &= 2\xi_{12}\xi_{1456} + (\xi_4+\xi_5+\xi_6)^2 - \xi_4^2 + \xi_5^2 - \xi_6^2,\end{aligned}$$

which implies that

$$\left|\sum_{i=1}^{6}(-1)^{i-1}\xi_i^2\right| \gtrsim N^2,$$

and for $\lambda_1 + \cdots + \lambda_6 = 0$

$$N^2 \lesssim \max_{i=1,\dots,6} |\lambda_i + (-1)^i \xi_i^2|. \tag{46}$$

If the integral in time were performed on the whole real line instead of $[T, T+\delta]$, then, after paying the price of the extra factor $\max_{i=1,\dots,6} |\lambda_i + (-1)^i\xi_i^2|^{\frac{1}{4}}$, one would obtain

$$\left|\int_T^{T+\delta}\int \Lambda_6(M_6(\xi_1,\xi_2,\dots,\xi_6))\,dt\right| \lesssim \frac{1}{N^{2-}}\|Iw\|^6_{1,\frac{3}{4}-}.$$

This argument has to be modified when the time integral is performed on a finite interval $[T, T+\delta]$, due to the fact that $\chi_{[T,T+\delta]}$, the characteristic function of the interval $[T, T+\delta]$, is not smooth enough. A similar difficulty was encountered also in [9]. We split

$$\chi_{[T,T+\delta]}(t) = a(t) + b(t),$$

where

$$\hat{a}(\tau) = \widehat{\chi_{[T,T+\delta]}}(\tau)\eta(\tau/N^2),$$

and η is supported on a small interval of 0 and equals 1 near 0, so a is smoothing out $\chi_{[T,T+\delta]}$ at scale N^{-2}. If one replaces $\chi_{[T,T+\delta]}(t)$ with $a(t)$, then the argument above works because the Fourier transform of $a(t)$ is supported on $|\tau| \ll N^2$, and one can still obtain the crucial inequality (46). We now have to deal with $b(t)$. It is easy to check that

$$\|b(t)\|_{L^1_t} \lesssim N^{-2}.$$

So we just have to show that

$$\sup_t |\Lambda_6(M_6; w_1(t),\dots,w_6(t))| \lesssim \prod_{j=1}^{6}\|Iw_j\|_{X^{1,\frac{3}{4}-}}. \tag{47}$$

We can crudely use Lemma 6.4 and obtain

$$|\Lambda_6(M_6; w_1(t), \dots, w_6(t))| \lesssim m_{soprano}^2 N_{soprano}^2 \|w_{soprano}\|_{L_t^\infty L_x^2} \|w_{alto}\|_{L_t^\infty L_x^2} \times \|w_{tenor}\|_{L_t^\infty L_x^\infty} \|w_{baritone}\|_{L_t^\infty L_x^\infty} \prod_{j \notin S} \|Iw_j\|_{L_t^\infty L_x^\infty},$$

which gives (47) by the Sobolev embedding theorem.

- *Case* (c). $N_{soprano} \sim N_{alto} \gtrsim N \gg N_{tenor}$. By Lemma 6.4, Lemma 2.2, Sobolev inequality, and (45), we have

$$\begin{aligned} \left| \int\int \Lambda_6(M_6(\xi_1, \xi_2, \dots, \xi_6)) \right| &\lesssim \sum_{S,j} \frac{1}{m_{alto}^2 N_{alto}} N_{soprano} N_{tenor} \\ &\quad \times \|Iw_{soprano} Iw_{tenor}\|_{L^2} \\ &\quad \times N_{alto} \|Iw_{alto} Iw_{baritone}\|_{L^2} \prod_{j \notin S} \|w_j\|_{L^\infty} \\ &\lesssim \frac{1}{N^{2-}} \|Iw\|_{1, \frac{1}{2}+}. \end{aligned}$$

This concludes the proof of the lemma. □

REFERENCES

[1] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform,* SIAM Stud. Appl. Math. 4, SIAM, Philadelphia, 1981.

[2] H. Biagioni and F. Linares, *Ill-posedness for the derivative Schrödinger and generalized Benjamin-Ono equations,* Trans. Amer. Math. Soc., 353 (2001), pp. 3649–3659.

[3] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations, Part* I, Geom. Funct. Anal., 3 (1993), pp. 107–156.

[4] J. Bourgain, *Refinements of Strichartz' inequality and applications to* 2*D-NLS with critical nonlinearity,* Internat. Math. Res. Notices, 5 (1998), pp. 253–283.

[5] J. Bourgain, *Periodic Korteweg-de Vries equation with measures as initial data,* Selecta Math. (N.S.), 3 (1997), pp. 115–159.

[6] J. Bourgain, *New Global Well-Posedness Results for Nonlinear Schrödinger Equations,* AMS, Providence, RI, 1999.

[7] T. Cazenave and F. Weissler, *The Cauchy problem for the critical nonlinear Schrödinger equation in* H^s*,* Nonlinear Anal., 14 (1990), pp. 807–836.

[8] P. Clarkson and C. Cosgrove, *Painlevé analysis of the nonlinear Schrödinger family of equations,* J. Phys. A, 20 (1987), pp. 2003–2024.

[9] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Global well-posedness for Schrödinger equations with derivative,* SIAM J. Math. Anal., 33 (2001), pp. 649–669.

[10] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Almost Conservation Laws and Global Rough Solutions to a Nonlinear Schrödinger Equation,* preprint.

[11] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Sharp Global Well-Posedness for Periodic and Non-Periodic KdV and mKdV,* preprint.

[12] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Multilinear Estimates for Periodic KdV Equations and Applications,* preprint.

[13] J. Ginibre, *The Cauchy problem for periodic semilinear PDE in space variables (after Bourgain),* Astérisque, 237 (1996), pp. 163–187.

[14] N. Hayashi, *The initial value problem for the derivative nonlinear Schrödinger equation in the energy space,* Nonlinear Anal., 20 (1993), pp. 823–833.

[15] N. Hayashi and T. Ozawa, *On the derivative nonlinear Schrödinger equation,* Phys. D, 55 (1992), pp. 14–36.

[16] N. Hayashi and T. Ozawa, *Finite energy solution of nonlinear Schrödinger equations of derivative type,* SIAM J. Math. Anal., 25 (1994), pp. 1488–1503.

[17] N. Hayashi and T. Ozawa, *Remarks on nonlinear Schrödinger equations in one space dimension,* Differential Integral Equations, 2 (1994), pp. 453–461.

[18] N. Hayashi and T. Ozawa, *Modified wave operators for the derivative nonlinear Schródinger equation,* Math. Ann., 298 (1994), pp. 557–576.
[19] D. J. Kaup and A. C. Newell, *An exact solution for a derivative Schródinger equation,* J. Math. Phys., 19 (1978), pp. 798–801.
[20] M. Keel and T. Tao, *Local and global well-posedness of wave maps on $\mathbb{R}^{1+1}$ for rough data,* Internat. Math. Res. Notices, 21 (1998), pp. 1117–1156.
[21] M. Keel and T. Tao, *Global Well-Posedness of the Maxwell-Klein-Gordon Equation Below the Energy Norm,* preprint.
[22] C. E. Kenig, G. Ponce, and L. Vega, *The Cauchy problem for the Korteweg-de Vries equation in Sobolev spaces of negative indices,* Duke Math. J., 71 (1993), pp. 1–21.
[23] C. Kenig, G. Ponce, and L. Vega, *A bilinear estimate with applications to the KdV equation,* J. Amer. Math. Soc., 9 (1996), pp. 573–603.
[24] C. Kenig, G. Ponce, and L. Vega, *Small solutions to nonlinear Schrödinger equations,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 255–288.
[25] W. Mio, T. Ogino, K. Minami, and S. Takeda, *Modified nonlinear Schrödinger for Alfvén waves propagating along the magnetic field in cold plasma,* J. Phys. Soc. Japan, 41 (1976), pp. 265–271.
[26] E. Mjolhus, *On the modulational instability of hydromagnetic waves parallel to the magnetic field,* J. Plasma. Phys., 16 (1996), pp. 321–334.
[27] T. Ozawa, *On the nonlinear Schrödinger equations of derivative type,* Indiana Univ. Math. J., 45 (1996), pp. 137–163.
[28] T. Ozawa and Y. Tsutsumi, *Space-time estimates for null gauge forms and nonlinear Schrödinger equations,* Differential Integral Equations, 11 (1998), pp. 201–222.
[29] C. Sulem and P.-L. Sulem, *The Nonlinear Schrödinger Equation,* Appl. Math. Sci. 139, Springer-Verlag, New York, 1999.
[30] H. Takaoka, *Well-posedness for the one dimensional Schrödinger equation with the derivative nonlinearity,* Adv. Differential Equations, 4 (1999), pp. 561–680.
[31] H. Takaoka, *Global well-posedness for Schrödinger equations with derivative in a nonlinear term and data in low-order Sobolev spaces,* Electron. J. Differential Equations, 42 (2001), pp. 1–23.
[32] T. Tao, *Multilinear weighted convolution of L^2 functions, and applications to nonlinear dispersive equations,* Amer. J. Math., 123 (2001), pp. 839–908.
[33] Y. Tsutsumi, *L^2 solutions for nonlinear Schrodinger equations and nonlinear groups,* Funkcial. Ekvac., 30 (1987), pp. 115–125.
[34] M. Tsutsumi and I. Fukuda, *On solutions of the derivative nonlinear Schrödinger equation: Existence and uniqueness theorem,* Funkcial. Ekvac., 23 (1980), pp. 259–277.
[35] M. Tsutsumi and I. Fukuda, *On solutions of the derivative nonlinear Schrödinger equation* II, Funkcial. Ekvac., 234 (1981), pp. 85–94.
[36] M.I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates,* Comm. Math. Phys., 87 (1983), pp. 567–576.

SIAM J. MATH. ANAL.
Vol. 34, No. 1, pp. 87–100

ON THE HÖLDER CONTINUITY OF SOLUTIONS OF A CERTAIN SYSTEM RELATED TO MAXWELL'S EQUATIONS*

KYUNGKEUN KANG† AND SEICK KIM†

Abstract. We study the system $\operatorname{curl}(a(x)\operatorname{curl} u) = 0$, $\operatorname{div} u = 0$ with a bounded measurable coefficient $a(x)$. The main result of this paper is the Hölder continuity of weak solutions of the system above. As an application, we prove the C^α regularity of weak solutions of the Maxwell's equations in a quasi-stationary electromagnetic field.

Key words. interior regularity, Maxwell's equations, measurable coefficients

AMS subject classifications. 35B45, 35J60, 35Q60

PII. S0036141001393341

1. Introduction.

Let Ω be a domain in $\mathbb{R}^3$ and $a \in L^\infty(\Omega)$ be a scalar function bounded by two positive numbers m, M. In this paper we study the regularity problem of the following system:

$$\left.\begin{array}{c} \nabla \times [a(x)\nabla \times u] = f + \nabla \times g \\ \nabla \cdot u = 0 \end{array}\right\} \quad \text{in} \quad \Omega. \tag{1.1}$$

Here we denote $\nabla \times u = \operatorname{curl} u$ and $\nabla \cdot u = \operatorname{div} u$. The question about the regularity of the solution of such a system was raised by Professor M. Giaquinta. The main result of this paper is the Hölder continuity of weak solutions of system (1.1) under appropriate assumptions on the inhomogeneous terms f, g.

The above system arises from Maxwell's equations in a quasi-stationary electromagnetic field where the displacement of electrical current is assumed to be time independent. We are grateful to Professor M. Hong for valuable discussions elucidating the connection between the system (1.1) and Maxwell's equations. In the study of the penetration of a magnetic field in materials, the electrical resistance strongly depends on the temperature, and, by taking the temperature effect into consideration, the classical Maxwell system in a quasi-stationary electromagnetic field reduces to the following mathematical model: find $H(x,t)$ and $u(x,t)$ such that

$$\begin{cases} H_t + \nabla \times [\sigma(u)\nabla \times H] = 0, \\ \nabla \cdot H = 0, \\ u_t - \Delta u = \sigma(u)\,|\nabla \times H|^2, \end{cases} \tag{1.2}$$

where H and u represent, respectively, the strength of the magnetic field and temperature, while $\sigma(u)$ denotes the electrical resistivity of the material (see, e.g., [9], [10]). In particular, in the steady state we have the following steady-state system:

$$\begin{cases} \nabla \times [\sigma(u)\nabla \times H] = 0, \\ \nabla \cdot H = 0, \\ -\Delta u = \sigma(u)\,|\nabla \times H|^2. \end{cases} \tag{1.3}$$

*Received by the editors August 2, 2001; accepted for publication (in revised form) March 6, 2002; published electronically August 15, 2002.
http://www.siam.org/journals/sima/34-1/39334.html
†School of Mathematics, University of Minnesota, 206 Church Street S.E., Minneapolis, MN 55455 (kkang@math.umn.edu, skim@math.umn.edu). The first author was supported in part by NSF grant DMS-9877055. The second author was supported in part by NSF grant DMS-9971052.

Global existence of a pair of weak solutions (H, u) of the system (1.2) was established by Yin [9]. However, the continuity of weak solutions of the system (1.2) as well as the system (1.3) was unknown. In section 3, we will show that, by using our result on the linear system (1.1), weak solutions of the coupled nonlinear system (1.3) are locally Hölder continuous.

As we mentioned earlier, the motivation for studying the system (1.1) is an interesting question which has been raised by Giaquinta and Hong [4]. The original formulation of the question appears in terms of differential forms. However, in the case when $n = 3$, it can be rephrased as follows: Are weak solutions to the following system locally Hölder continuous?

$$\left.\begin{array}{c} \nabla \times [a(x)\nabla \times u] = 0 \\ \nabla \cdot u = 0 \end{array}\right\} \quad \text{in} \quad \Omega. \tag{1.4}$$

Indeed, the above system (1.4) is a special case of (1.1), and, as we mentioned at the beginning, the answer is positive when $n = 3$. We don't know the answer for higher dimensions $n \geq 4$. In section 4 we will formulate their original question by using differential forms and discuss some related problems. Very recently, we received a preprint by Yin [11], in which a similar result to ours is obtained. It seems [11] used an idea similar to ours, although technical details are different.

2. Main results: Hölder estimates. In this section we shall always assume $n = 3$. First, we will introduce notations.

- For $x \in \mathbb{R}^n$ and $\rho > 0$, we define $B_\rho(x) = \{y \in \mathbb{R}^n : |x - y| < \rho\}$.
- For a measurable set $S \subset \mathbb{R}^n$, we define $\fint_S f = \frac{1}{|S|}\int_S f\,dx$.
- We denote $(f)_{x,\rho} = f_{x,\rho} = \fint_{B_\rho(x)} f\,dx$.
- We denote $B_\rho = B_\rho(x)$ and $f_\rho = f_{x,\rho}$ if x is clear from the context.
- Let $\mathcal{D}(\Omega) = \mathcal{D}(\Omega; \mathbb{R}^n) = \{f \in C^\infty(\Omega; \mathbb{R}^n) : \nabla \cdot f = 0\}$. We denote by $\mathcal{H}^q(\Omega)$, $q \in [1, \infty)$, the completion of $\mathcal{D}(\Omega)$ in the norm of $L^q(\Omega)$.
- $\Omega' \Subset \Omega$ means Ω' is a precompact subset of Ω.
- For $u = (u^1, \dots, u^n)$, we denote by ∇u the gradient matrix: $(\nabla u)_{ij} = D_j u^i$.

Now we will state our main results. Consider the following linear system:

$$\left.\begin{array}{c} \nabla \times [a(x)\nabla \times u] = f \\ \nabla \cdot u = 0 \end{array}\right\} \quad \text{in} \quad \Omega, \tag{2.1}$$

where $f \in \mathcal{H}^q(\Omega; \mathbb{R}^n)$ and $a \in L^\infty(\Omega; \mathbb{R})$ such that $m \leq a \leq M$ for some constants $m, M > 0$. The restriction $f \in \mathcal{H}^q(\Omega; \mathbb{R}^n)$ arises from the consistency condition

$$0 = \nabla \cdot \nabla \times [a(x)\nabla \times u] = \nabla \cdot f$$

in the sense of distribution.

THEOREM 2.1. *Let $u \in W^{1,2}_{loc}(\Omega; \mathbb{R}^n)$ be a weak solution of the system (2.1) with $f \in \mathcal{H}^q_{loc}(\Omega)$, $q > n/2$. Let $B := B_R(x_0) \Subset \Omega$. Then there exist constants $\alpha = \alpha(m, M, q) > 0$ and $C = C(m, M, q, R)$ such that u is Hölder continuous in $B_{R/16}(x_0)$ and*

$$\|u\|_{C^{0,\alpha}(\overline{B}_{R/16})} \leq C\left[\|u\|_{L^2(B)} + \|f\|_{L^q(B)}\right]. \tag{2.2}$$

Next we consider the quasi-linear system

$$\left.\begin{array}{c}\nabla\times[\sigma(x,u)\nabla\times u]=f\\ \nabla\cdot u=0\end{array}\right\}\quad\text{in}\quad\Omega. \tag{2.3}$$

Here we assume $f\in\mathcal{H}^q_{loc}(\Omega)$ and $\sigma:\Omega\times\mathbb{R}^n\to\mathbb{R}$ satisfies the following conditions:

(a) $m\le\sigma\le M$ for positive constants m,M.

(b) σ is Hölder continuous in $\Omega\times\mathbb{R}^n$: $[\sigma]_\mu=[\sigma]_{C^{0,\mu}(\Omega\times\mathbb{R}^n)}<\infty$ for some $\mu\in(0,1)$.

THEOREM 2.2. *Let $u\in W^{1,2}_{loc}(\Omega;\mathbb{R}^n)$ be a weak solution of (2.3) and assume $f\in\mathcal{H}^p_{loc}(\Omega)$, $p>n$. Let $B:=B_R(x_0)\Subset\Omega$. Then, under the above assumptions on σ, ∇u is locally Hölder continuous with exponent $\alpha=\min(\mu,1-n/p)$ and*

$$\|u\|_{C^{1,\alpha}(\overline{B}_{R/4})}\le C,\quad C=C(m,M,p,[\sigma]_\mu,\|u\|_{W^{1,2}(B)},\|f\|_{L^p(B)},R).$$

The following technical lemmas will be used in the proof of theorems.

LEMMA 2.3 (uniqueness). *Let $u\in W^{1,2}_0(\Omega)$ be a weak solution of*

$$\left.\begin{array}{c}\nabla\times[a(x)\nabla\times u]=0\\ \nabla\cdot u=0\end{array}\right\}\quad in\quad\Omega.$$

Then $u\equiv0$ in Ω.

Proof. Since $\nabla\cdot u=0$, integration by parts yields

$$\int_\Omega|\nabla\times u|^2=\int_\Omega|\nabla\times u|^2+|\nabla\cdot u|^2=\int_\Omega|\nabla u|^2.$$

On the other hand, by using u itself as a test function we have

$$m\int_\Omega|\nabla\times u|^2\le\int_\Omega a(x)\,|\nabla\times u|^2=\int_\Omega\nabla\times[a(x)\,\nabla\times u]\cdot u=0.$$

Hence, $\nabla u=0$ in Ω. This completes the proof. □

LEMMA 2.4. *Let $B\subset\mathbb{R}^n$ be a open ball and let $f\in\mathcal{D}(B)$. Then there exists $g\in C^\infty(B;\mathbb{R}^n)\cap\mathcal{D}(B)$ such that $\nabla\times g=f$ in B and $g=0$ on ∂B. Moreover, if $f\in\mathcal{H}^p(B)$, $1<p<\infty$, then $\|\nabla g\|_{L^p(B)}\le C(p)\,\|f\|_{L^p(B)}$.*

Proof. Let g be the unique solution of

$$\left\{\begin{array}{rl}-\Delta g=\nabla\times f & \text{in}\quad B,\\ g=0 & \text{on}\quad\partial B.\end{array}\right.$$

From the following vector identity,

$$\nabla\times(\nabla\times f)=\nabla(\nabla\cdot f)-\Delta f, \tag{2.4}$$

and the representation formula of g in terms of the Green's function, it is easy to see $\nabla\times g=f$ and $\nabla\cdot g=0$ in B. The second part of the lemma follows from the L^p theory of the Laplace operator. □

LEMMA 2.5. *Suppose $F\in C^\infty(B;\mathbb{R}^n)$ satisfies $\nabla\times F=0$ in B. Then there exists $\varphi\in C^\infty(B;\mathbb{R})$ such that $\nabla\varphi=F$ in B and $\varphi=0$ on ∂B. Moreover, if $F\in L^2(B)$, then $\|\varphi\|_{L^2}\le C\,\|F\|_{L^2}$.*

Proof. Let φ be the unique solution of

$$\left\{\begin{array}{rl}\Delta\varphi=\nabla\cdot F & \text{in}\quad B,\\ \varphi=0 & \text{on}\quad\partial B.\end{array}\right.$$

Then $\nabla\varphi = F$ will follow immediately from Lemma 2.3. Also, $\|\nabla\varphi\|_{L^2} \le C\|F\|_{L^2}$. Since $\varphi = 0$ on ∂B, we can use Poincaré inequality to get $\|\varphi\|_{L^2(B)} \le C\|F\|_{L^2}$. □

LEMMA 2.6. *Let $w \in W_0^{1,2}(B_R;\mathbb{R}^n)$ be a weak solution of*

$$\left.\begin{array}{c}\nabla\times(\nabla\times w) = \nabla\times(F+G+H)\\ \nabla\cdot w = 0\end{array}\right\} \quad \text{in} \quad B_R,$$

where $F \in C^{0,\mu}(B_R)$, $\mu > 0$, $G \in L^2(B_R)$, and $H \in L^q(B_R)$, $q > n$. Then

$$\int_{B_R} |\nabla w|^2 \le C\left([F]_\mu^2 R^{n+2\mu} + \|G\|_{L^2}^2 + \|H\|_{L^q}^2 R^{n-2+2\gamma}\right), \qquad \gamma = 1 - n/q > 0.$$

Proof. From the identity (2.4), $w \in W_0^{1,2}$ is a weak solution of

$$-\Delta w = \nabla\times(F+G+H) \quad \text{in} \quad B_R.$$

By using w itself as a test function we get

$$\int_{B_R} \nabla w\cdot\nabla w = \int_{B_R} (F-F_R)\cdot\nabla\times w + \int_{B_R}(G+H)\cdot\nabla\times w.$$

Hence, Schwarz inequality yields

$$\int_{B_R}|\nabla w|^2 \le \int_{B_R}|F-F_R|^2 + \int_{B_R}|G|^2 + \int_{B_R}|H|^2 + \frac{3}{4}\int_{B_R}|\nabla\times w|^2.$$

Since $\nabla\cdot w = 0$, integration by parts yields

$$\int_{B_R}|\nabla\times w|^2 = \int_{B_R}|\nabla\times w|^2 + |\nabla\cdot w|^2 = \int_{B_R}|\nabla w|^2.$$

The lemma follows from obvious inequalities $\int_{B_R}|F-F_R|^2 \le C(n)[F]_\mu^2 R^{n+2\mu}$ and $\|H\|_{L^2}^2 \le \|H\|_{L^q}^2 |B_R|^{1-2/q}$. □

LEMMA 2.7. *Let $u \in W^{1,2}(B_2;\mathbb{R}^n)$ be a weak solution of*

$$\nabla\times[a(x)\nabla\times u] = \nabla\times g \quad \text{in} \quad \Omega.$$

Then $\|\nabla\times u\|_{L^2(B_1)} \le C(\|u\|_{L^2(B_2)} + \|g\|_{L^2(B_2)})$.

Proof. This is a Caccioppoli-type inequality. The proof is straightforward. □

LEMMA 2.8. *Let $\phi(t)$ be a nonnegative and nondecreasing function. Suppose that*

$$\phi(\rho) \le A\left[\left(\frac{\rho}{r}\right)^\alpha + \varepsilon\right]\phi(r) + Br^\beta$$

for all $\rho < r \le R_0$, with A, α, β nonnegative constants, $\beta < \alpha$. Then there exists a constant $\varepsilon_0 = \varepsilon_0(A,\alpha,\beta)$ such that, if $\varepsilon < \varepsilon_0$ for all $\rho < r \le R_0$, we have

$$\phi(\rho) \le c\left[\left(\frac{\rho}{r}\right)^\beta \phi(r) + B\rho^\beta\right],$$

where c is a constant depending on α, β, A.

Proof. See [3, Lemma 2.1, p. 86]. □

Now we are ready to prove our main theorems.

Proof of Theorem 2.1. First, we shall assume that $a \in L^\infty(\Omega) \cap C^\infty(\Omega)$ and $f \in \mathcal{H}^q_{loc}(\Omega) \cap C^\infty(\Omega)$. The constant C which appears in (2.2) will not depend on extra smoothness of data. Since (2.1) is a linear system, the full result will then follow from Lemma 2.3 and standard approximation argument. Also, we will assume without loss of generality that $B = B_{16}(x_0)$. Moreover, we may assume $n/2 < q < n$. The case $q \geq n$ will be recovered by Hölder's inequality.

Since $\nabla \cdot f = 0$, we conclude from Lemma 2.4 that there exists $g \in C^\infty(B_8; \mathbb{R}^n)$ such that $f = \nabla \times g$ and $g = 0$ on ∂B_8. Then Sobolev–Poincaré inequality implies

$$\|g\|_{L^{q^*}(B_8)} \leq C \|\nabla g\|_{L^q(B_8)} \leq C \|f\|_{L^q(B_8)}, \qquad q^* = nq/(n-q) > n. \tag{2.5}$$

Then by Lemma 2.5 there exists $\varphi \in C^\infty(B_8; \mathbb{R})$ such that

$$\nabla \varphi = a(x) \nabla \times u - g \tag{2.6}$$

and

$$\|\varphi\|_{L^2(B_8)} \leq C \Big(\|\nabla \times u\|_{L^2(B_8)} + \|g\|_{L^2(B_8)} \Big). \tag{2.7}$$

From Lemma 2.7 and (2.5), we can estimate $\|\varphi\|_{L^2(B_8)}$ in (2.7):

$$\|\varphi\|_{L^2(B_8)} \leq C \Big(\|u\|_{L^2(B)} + \|f\|_{L^q(B)} \Big). \tag{2.8}$$

By rewriting (2.6) as $\nabla \times u = a^{-1} \nabla \varphi + a^{-1} g$ we conclude

$$0 = \nabla \cdot (\nabla \times u) = \nabla \cdot [a^{-1} \nabla \varphi] + \nabla \cdot (a^{-1} g).$$

Now we have a single elliptic equation

$$-\nabla \cdot [a^{-1} \nabla \varphi] = \nabla \cdot (a^{-1} g). \tag{2.9}$$

It is well known that the following estimate holds:

$$\|\varphi\|_{C^{0,\beta}(B_4)} \leq C \Big(\|\varphi\|_{L^2(B_8)} + \|g\|_{L^{q^*}(B_8)} \Big), \tag{2.10}$$

where $C = C(M/m, q)$ and $\beta = \beta(M/m) > 0$ (see, e.g., [5, Theorem 8.24]).

Also, from (2.9) we have the following Caccioppoli inequality: for all $r \leq 4$

$$\begin{aligned} \int_{B_{r/2}} |\nabla \varphi|^2 &\leq C \left(\frac{1}{r^2} \int_{B_r} |\varphi - (\varphi)_r|^2 + \int_{B_r} |g|^2 \right) \\ &\leq C \Big([\varphi]^2_{C^{0,\beta}(B_4)} r^{n-2+2\beta} + \|g\|^2_{L^{q^*}(B_8)}\, r^{n-2+2\gamma} \Big), \end{aligned} \tag{2.11}$$

where $C = C(M/m)$ and $\gamma = (2 - \frac{n}{q}) > 0$.

Since $\nabla \cdot u = 0$, (2.4) implies

$$-\Delta u = \nabla \times (\nabla \times u) = \nabla \times (a^{-1} \nabla \varphi) + \nabla \times (a^{-1} g).$$

Fix $r \leq 2$ and decompose u into two functions v and $w := u - v$ such that v is the unique solution of

$$\begin{cases} -\Delta v = 0 & \text{in} \quad B_r, \\ v = u & \text{on} \quad \partial B_r. \end{cases}$$

Then $w = 0$ on ∂B_r and solves

$$-\Delta w = \nabla \times (a^{-1}\nabla\varphi) + \nabla \times (a^{-1}g) \quad \text{in} \quad B_r.$$

Hence, from (2.11) and Lemma 2.6 (with $G = a^{-1}\nabla\varphi$ and $H = a^{-1}g$), together with Poincaré inequality, we get

$$\int_{B_r} |w - w_r|^2 \le C \left([\varphi]^2_{C^{0,\beta}(B_4)} r^{n+2\beta} + \|g\|^2_{L^{q^*}(B_8)} r^{n+2\gamma} \right).$$

Then, since v is harmonic, the following estimates hold for all $\rho < r \le 2$:

$$\begin{aligned} \int_{B_\rho} |u - u_\rho|^2 &\le C \left(\frac{\rho}{r}\right)^{n+2} \int_{B_r} |u - u_r|^2 + C \int_{B_r} |w - w_r|^2 \\ &\le C \left(\frac{\rho}{r}\right)^{n+2} \int_{B_r} |u - u_r|^2 \\ &\quad + C \left([\varphi]^2_{C^{0,\beta}(B_4)} r^{n+2\beta} + \|g\|^2_{L^{q^*}(B_8)} r^{n+2\gamma} \right), \end{aligned} \tag{2.12}$$

where $C = C(m, M)$.

Let $\phi(\rho) := \int_{B_\rho} |u - u_\rho|^2$ and $\alpha = \min(\beta, \gamma)$. Combining (2.5), (2.8), and (2.10),

$$\phi(\rho) \le C \left[\left(\frac{\rho}{r}\right)^{n+2} \phi(r) + r^{n+2\alpha} \left(\|u\|^2_{L^2(B)} + \|f\|_{L^q(B)} \right) \right]. \tag{2.13}$$

Since (2.13) holds for any $\rho < r \le 2$, by Campanato's integral characterization of Hölder continuous function, together with Lemma 2.8, we conclude that

$$[u]_{C^{0,\alpha}(B_2)} \le C(m, M, q) \left(\|u\|_{L^2(B)} + \|f\|_{L^q(B)} \right). \tag{2.14}$$

Fix $x \in \overline{B}_1$ and consider a ball $B_1(x) \subset B_2$. Then

$$|u(x)| \le |u(y)| + |u(x) - u(y)| \le |u(y)| + [u]_{C^{0,\alpha}(B_2)} \quad \forall y \in B_1(x). \tag{2.15}$$

Integrating (2.15) with respect to y over $B_1(x)$ we get

$$|u(x)| \le C \left(\|u\|_{L^2(B_2)} + [u]_{C^{0,\alpha}(B_2)} \right) \quad \forall x \in B_1. \tag{2.16}$$

Combining (2.14) and (2.16) we finally obtain

$$\|u\|_{C^{0,\alpha}(\overline{B}_1)} \le C(m, M, q) \left(\|u\|_{L^2(B)} + \|f\|_{L^q(B)} \right).$$

This completes the proof. □

For the proof of Theorem 2.2, we need $C^{1,\alpha}$ estimates of the linear system (2.1) under the assumption that a is Hölder continuous.

LEMMA 2.9. *Let $u \in W^{1,2}_{loc}(\Omega; \mathbb{R}^n)$ be a weak solution of (2.1) where $f \in \mathcal{H}^p_{loc}(\Omega)$ for some $p > n$. Assume further that $a \in C^{0,\mu}(\Omega; \mathbb{R})$. Then, if $B := B_R(x_0) \Subset \Omega$, ∇u is Hölder continuous in $B_{R/4}(x_0)$ and*

$$[\nabla u]_{C^{0,\alpha}(B_{R/4})} \le C \left(\|\nabla u\|_{L^2(B)} + \|f\|_{L^q(B)} \right). \tag{2.17}$$

Here, $\alpha = \min(1 - n/p, \mu)$ *and* $C = C(n, m, M, p, [a]_\mu, R)$.

Proof. The proof relies on the standard perturbation method. As in Theorem 2.1, we may assume that f is smooth and $B = B_4(x_0)$. Then, by Lemma 2.4 there exists g such that $f = \nabla \times g$ and $[g]_{C^{0,\nu}(B)} \le C \|\nabla g\|_{L^p(B)} \le C \|f\|_{L^p(B)}$, $\nu = 1 - n/p > 0$. Let $y \in \overline{B}_2(0)$ and let $R_0 \le 2$ be a fixed number which will be specified later. Then

$$a(y)[\nabla \times (\nabla \times u)] = \nabla \times (\,[a(y) - a(x)]\ \nabla \times u) + \nabla \times g \quad \text{in} \quad B_{R_0}(y) \subset B.$$

Fix an $r \le R_0$ and split u into v and $w := u - v$ such that

$$\begin{cases} -\Delta v = 0 & \text{in} \quad B_r(y), \\ v = u & \text{on} \quad \partial B_r(y). \end{cases}$$

Then $w \in W_0^{1,2}(B_r(y))$ and satisfies

$$-a(y)\Delta w = \nabla \times (\,[a(y) - a(x)]\,\nabla \times u + g) \quad \text{in} \quad B_r(y).$$

Hence, from Lemma 2.6 with $F = g$ and $G = [a(y) - a(x)]\nabla \times u$, we obtain

$$\begin{aligned} \int_{B_r(y)} |\nabla w|^2 &\le C \left([a]_\mu^2 r^{2\mu} \|\nabla \times u\|^2_{L^2(B_r(y))} + [g]_\nu^2\, r^{n+2\nu} \right) \\ &\le C \left(r^{2\mu} \int_{B_r(y)} |\nabla u|^2 + \|f\|^2_{L^p(B)}\ r^{n+2\nu} \right). \end{aligned}$$

Since ∇v is harmonic in $B_r(y)$, the following estimate holds for $\rho < r \le R_0$:

$$\begin{aligned} \int_{B_\rho(y)} |\nabla u|^2 &\le C \left[\left(\frac{\rho}{r}\right)^n \int_{B_r(y)} |\nabla u|^2 + \int_{B_r(y)} |\nabla w|^2 \right] \\ &\le C \left[\left(\frac{\rho}{r}\right)^n + r^{2\mu} \right] \int_{B_r(y)} |\nabla u|^2 + C \|f\|^2_{L^p(B)}\ r^{n+2\nu}. \end{aligned}$$

We will apply Lemma 2.8 to the quantity $\phi(\rho) := \int_{B_\rho(y)} |\nabla u|^2$. Choose R_0 small enough so that $R_0^{2\mu} < \varepsilon_0$. Then Lemma 2.8 implies

$$\int_{B_\rho(y)} |\nabla u|^2 \le c\rho^{n-\mu} \left[\|\nabla u\|^2_{L^2(B)} + \|f\|^2_{L^p(B)} \right] \quad \forall y \in \overline{B}_2, \quad \forall \rho \le R_0.$$

Now set $y = x_0$ and $R_0 = 2$. In the rest of the proof we will denote $B_r := B_r(x_0)$. By using standard covering argument, if necessary, we obtain

$$\int_{B_r} |\nabla u|^2 \le C r^{n-\mu} \left[\|\nabla u\|^2_{L^2(B)} + \|f\|^2_{L^p(B)} \right] \quad \forall r \le 2. \tag{2.18}$$

On the other hand, for all $\rho < r \le 2$,

$$\begin{aligned} \int_{B_\rho} |\nabla u - (\nabla u)_\rho|^2 &\le C \left[\left(\frac{\rho}{r}\right)^{n+2} \int_{B_r} |\nabla u - (\nabla u)_r|^2 + \int_{B_r} |\nabla w|^2 \right] \\ &\le C \left[\left(\frac{\rho}{r}\right)^{n+2} \int_{B_r} |\nabla u - (\nabla u)_r|^2 \right] \\ &\quad + C \left(\|\nabla u\|^2_{L^2(B_r)}\, r^{2\mu} + \|f\|^2_{L^p(B)}\, r^{n+2\nu} \right). \end{aligned} \tag{2.19}$$

Combining (2.18) and (2.19) we conclude that $\nabla u \in C^{0,\gamma}(B_1)$, $\gamma = \min(\nu, \mu/2)$. In particular, as in (2.16) in the proof of Theorem 2.1,

$$\sup_{B_1} |\nabla u| \le C \left(\|\nabla u\|_{L^2(B_2)} + \|f\|_{L^p(B)} \right). \tag{2.20}$$

We may then use inequality (2.19) again for $\rho < r \le 1$, getting

$$[\nabla u]_{C^{0,\alpha}(B_1)} \le C \left(\|\nabla u\|_{L^2(B)} + \|f\|_{L^p(B)} \right), \quad \alpha = \min(\mu, \nu).$$

This completes the proof. □

Proof of Theorem 2.2. First, by Theorem 2.1 we know $u \in C^{0,\beta}_{loc}(\Omega)$ for some $\beta > 0$. Then $a(x) := \sigma(x, u(x))$ is locally Hölder continuous with some exponent $\gamma > 0$. Hence, from Lemma 2.9 we conclude ∇u is locally Hölder continuous. In particular, ∇u is bounded in B. As in (2.20) we have an estimate

$$\sup_{B_{R/2}} |\nabla u| \le C \left(\|\nabla u\|_{L^2(B)} + \|f\|_{L^p(B)} \right).$$

Thus $a(x)$ is Hölder continuous in $B_{R/2}$ with exponent μ and $[a]_{C^{0,\mu}(B_{R/2})} \le K$, where K is a constant that depends on $\|\nabla u\|_{L^2(B)}$, $\|f\|_{L^p(B)}$, $[\sigma]_\mu$, and other prescribed quantities independent of u, f. Now the theorem follows from Lemma 2.9. □

Remark 2.10. *In the proof of Theorem* 2.1 *we actually proved that if* $f \in \mathcal{H}^{q/2}_{loc}(\Omega; \mathbb{R}^n)$ *and* $g \in L^q_{loc}(\Omega; \mathbb{R}^n)$, $q > n$, *then any weak solution of the system*

$$\left.\begin{array}{c} \nabla \times [a(x) \nabla \times u] = f + \nabla \times g \\ \nabla \cdot u = 0 \end{array}\right\} \quad \text{in} \quad \Omega \tag{2.21}$$

satisfies the following estimate in $B := B_R(x_0) \Subset \Omega$:

$$\|u\|_{C^{0,\alpha}(\overline{B}_{R/2})} \le C \left[\|u\|_{L^2(B)} + \|f\|_{L^{q/2}(B)} + \|g\|_{L^q(B)} \right]. \tag{2.22}$$

Also, the proof of Lemma 2.9 *implies that a weak solution of*

$$\left.\begin{array}{c} \nabla \times [\sigma(x,u) \nabla \times u] = f + \nabla \times g \\ \nabla \cdot u = 0 \end{array}\right\} \quad \text{in} \quad \Omega, \tag{2.23}$$

where $f \in \mathcal{H}^p_{loc}(\Omega)$, $p > n$, *and* $g \in C^{0,\beta}_{loc}(\Omega)$, $\beta > 0$, *is locally Hölder continuous with exponent* $\alpha = \min(\mu, 1 - n/p, \beta)$.

Remark 2.11. *In the two-dimensional case, the Hölder continuity of weak solutions of* (2.1) *may follow from Sobolev imbedding. In fact, if* $f \equiv 0$, *then a weak solution* u *belongs to* $W^{1,p}_{loc}(\Omega)$ *for all* $p \in (1, \infty)$. *However, when* $n = 3$, $C^{0,\alpha}$ *regularity is the optimal result. To see this, consider a solution of the form* $u = (0, 0, u_3) : \Omega \to \mathbb{R}^3$. *Let us assume for simplicity that* $f \equiv 0$. *Then the system* (2.1) *becomes*

$$\begin{cases} D_3(a(x) D_1 u_3) = 0, \\ D_3(a(x) D_2 u_3) = 0, \\ D_1(a(x) D_1 u_3) + D_2(a(x) D_2 u_3) = 0, \\ D_3 u_3 = 0. \end{cases} \tag{2.24}$$

From the last equation of (2.24), *we can set* $v(x_1, x_2) := u_3(x_1, x_2, x_3)$. *It also follows that* $a(x)$ *depends only on* x_1 *and* x_2. *Then* v *solves the following equation of*

divergence form in two variables:

$$Lv := \sum_{i=1}^{2} D_i(a(x)D_i v) = 0 \quad in \quad \Omega.$$

The operator L is called an isotropic operator. Piccinini and Spagnolo showed that v is locally Hölder continuous with exponent $\alpha = \frac{4}{\pi} \arctan \sqrt{m/M}$ *(see* [7, Theorem 2, p. 396]*). To see that it is an optimal result, consult Example* 2 *of* [7] *on page* 400.

3. Application to a Maxwell system. As mentioned in the introduction, the problem we have analyzed so far arises from the Maxwell's system in a quasi-stationary electromagnetic field. Especially if the electric conductivity strongly depends on the temperature, then by taking the temperature effect into consideration the classical Maxwell system in a quasi-stationary electromagnetic field reduces to the following mathematical model (see [9, pp. 1029–1032]):

$$\begin{cases} H_t + \nabla \times (\sigma(u)\nabla \times H) = 0, \\ \nabla \cdot H = 0, \\ u_t - \Delta u = \sigma(u) \left|\nabla \times H\right|^2, \end{cases} \tag{3.1}$$

where H and u are unknowns representing, respectively, the strength of magnetic field and temperature, while $\sigma(u)$ denotes the electric resistivity of the material which is assumed to be strictly positive and bounded; i.e., there exist positive numbers m, M such that

$$0 < m \le \sigma(s) \le M \quad \forall s \in \mathbb{R}. \tag{3.2}$$

In [9], Yin proved, under appropriate assumptions on boundary and initial conditions, the existence of a pair of global weak solutions (H, u):

$$\begin{aligned} &H \in L^\infty\left(0, T; L^2(\Omega; \mathbb{R}^3)\right) \cap L^2\left(0, T; W^{1,2}(\Omega; \mathbb{R}^3)\right), \\ &u \in L^\infty\left(0, T; L^1(\Omega; \mathbb{R})\right) \cap L^q\left(0, T; W^{1,q}(\Omega; \mathbb{R})\right), \quad q \in [1, 5/4). \end{aligned}$$

In addition, he showed that if a pair of weak solutions (H, u) are continuous, then they are classical provided that σ is smooth enough. However, as pointed out by him, the continuity of weak solutions is unknown even if σ is smooth. Continuity of weak solutions of (3.1) heavily relies on the regularity theory of the following system with bounded measurable coefficient $a(x, t)$:

$$\left.\begin{array}{c} v_t + \nabla \times [a(x,t)\nabla \times v] = 0 \\ \nabla \cdot v = 0 \end{array}\right\} \quad \text{in} \quad Q, \tag{3.3}$$

where Q is the space-time cylinder $\Omega \times (0, T)$ for some $T > 0$. We don't know at this time whether or not weak solutions of the system (3.3) are Hölder continuous.

In this section, we consider instead the following fully steady-state systems introduced by Yin (see [9, p. 1031]):

$$\left.\begin{array}{c} \nabla \times (\sigma(u)\nabla \times H) = 0 \\ \nabla \cdot H = 0 \\ -\Delta u = \sigma(u) \left|\nabla \times H\right|^2 \end{array}\right\} \quad \text{in} \quad \Omega. \tag{3.4}$$

Using the results we obtained in previous section, we will show the $C^{0,\alpha}$ regularity of weak solutions of (3.4).

THEOREM 3.1. *Let (H,u) be a pair of weak solutions of (3.4). Then $(H,u) \in C^{0,\alpha}_{loc}(\Omega)$ for some $\alpha > 0$. Moreover, the following estimates hold in $\Omega' \Subset \Omega$:*

$$[H]_{C^{0,\alpha}(\Omega')} + [u]_{C^{0,\alpha}(\Omega')} \le C(m, M, \Omega', \Omega, \|H\|_{L^2}, \|u\|_{L^2}). \tag{3.5}$$

Proof. Let $B := B_{4R} = B_{4R}(x_0) \Subset \Omega$. We will show (u,H) is Hölder continuous in $B_R = B_R(x_0)$. Indeed, from the proof of Theorem 2.1 we have

$$\|H\|_{C^{0,\alpha}(B_{2R})} \le C(m, M, R) \|H\|_{L^2(B)}, \qquad \alpha = \alpha(m/M) > 0. \tag{3.6}$$

It remains to show that u is also Hölder continuous in B_R. Using a vector identity,

$$\nabla \cdot (F \times G) = (\nabla \times F) \cdot G - F \cdot (\nabla \times G), \tag{3.7}$$

together with the first equation $\nabla \times (\sigma(u) \nabla \times H) = 0$ of (3.4), we obtain

$$\nabla \cdot [H \times (\sigma(u) \nabla \times H)] = \sigma(u) |\nabla \times H|^2. \tag{3.8}$$

We rewrite the last equation of (3.4) as follows:

$$-\Delta u = \nabla \cdot [H \times (\sigma(u) \nabla \times H)]. \tag{3.9}$$

As before, fix $r \le R$ and split u into two parts v and $w := u - v$ such that

$$\begin{cases} -\Delta v = 0 & \text{in} \quad B_r, \\ v = u & \text{on} \quad \partial B_r. \end{cases}$$

Then, as in (2.12), the following estimate holds for $\rho < r \le R$:

$$\int_{B_\rho} |u - u_\rho|^2 \le C \left(\frac{\rho}{r}\right)^{n+2} \int_{B_r} |u - u_r|^2 + Cr^2 \int_{B_r} |\nabla w|^2. \tag{3.10}$$

We need to estimate $\|\nabla w\|^2_{L^2(B_r)}$. Since $w \in W^{1,2}_0(B_r)$ and satisfies

$$-\Delta w = \nabla \cdot [H \times (\sigma(u) \nabla \times H)] \quad \text{in} \quad B_r,$$

integration by parts and Schwarz inequality yields

$$\int_{B_r} |\nabla w|^2 \le 2 \int_{B_r} \sigma(u)^2 |H|^2 |\nabla \times H|^2 \le 2M^2 \int_{B_r} |H|^2 |\nabla \times H|^2. \tag{3.11}$$

Since H is continuous, it is bounded in B_r and thus from (3.11)

$$\int_{B_r} |\nabla w|^2 \le C \sup_{B_{2R}} |H|^2 \int_{B_r} |\nabla \times H|^2. \tag{3.12}$$

On the other hand, from the fact that H solves the first equation of (3.4) it follows that

$$\int_{B_r} |\nabla \times H|^2 \le \frac{C}{r^2} \int_{B_{2r}} |H - H_{2r}|^2 \le C [H]^2_{C^{0,\alpha}(B_{2R})} r^{n-2+2\alpha}. \tag{3.13}$$

Combining (3.12) and (3.13) together with (3.6) we obtain the required estimate

$$\|\nabla w\|^2_{L^2(B_r)} \le C(m, M, R) \|H\|^4_{L^2(B)} r^{n-2+2\alpha}. \tag{3.14}$$

Finally, by inserting (3.14) into (3.10) we conclude from Lemma 2.8

$$[u]_{C^{0,\alpha}(B_R)} \le C(m, M, R)\left(\|u\|_{L^2(B)} + \|H\|^2_{L^2(B)}\right). \tag{3.15}$$

Theorem 3.1 follows from (3.6), (3.15), and standard covering argument. □

THEOREM 3.2. *Let (H, u) be a pair of weak solutions of* (3.4). *Assume further that σ is Hölder continuous with exponent $\mu \in (0, 1)$. Then $H \in C^{1,\mu}_{loc}(\Omega)$ and $u \in C^{2,\mu}_{loc}(\Omega)$.*

Proof. First, by Theorem 2.1 we have $(u, H) \in C^{0,\alpha}_{loc}(\Omega)$, which in turn implies $\sigma(u)$ is Hölder continuous with exponent $\beta = \alpha\mu$. Then $H \in C^{1,\beta}_{loc}(\Omega)$ by Lemma 2.9 and thus $\sigma(u)\,|\nabla \times H|^2 \in C^{0,\beta}_{loc}(\Omega)$. Since u solves

$$-\Delta u = \sigma(u)\,|\nabla \times H|^2 \quad \text{in} \quad \Omega, \tag{3.16}$$

it follows from the theory of the Laplace operator that $u \in C^{2,\beta}_{loc}(\Omega)$. In particular, ∇u is locally bounded and thus $\sigma(u) \in C^{0,\mu}_{loc}(\Omega)$. By Lemma 2.9 again, $H \in C^{1,\mu}_{loc}(\Omega)$. Therefore $\sigma(u)\,|\nabla \times H|^2 \in C^{0,\mu}_{loc}(\Omega)$ and $u \in C^{2,\mu}_{loc}(\Omega)$ by (3.16). This completes the proof. □

REMARK 3.3. *Let (H, u) be a pair of weak solutions of* (3.4). *Suppose that $\sigma \in C^{k,\alpha}$, where k is a nonnegative integer and $0 < \alpha < 1$. Then*

$$H \in C^{k+1,\alpha}_{loc}(\Omega), \quad u \in C^{k+2,\alpha}_{loc}(\Omega). \tag{3.17}$$

In particular, if $\sigma \in C^{1,\alpha}$, then (H, u) is a pair of classical solutions.

4. Remarks on the case $n \ge 4$. First, we introduce some notations. Let Ω be a domain in $\mathbb{R}^n$, $n \ge 3$. Denote by $\Lambda^k := \Lambda^k(\Omega)$ the class of k-forms in Ω. Let $* : \Lambda^k \to \Lambda^{n-k}$ be the Hodge star linear operator, defined by setting

$$*(dx^{i_1} \wedge \cdots \wedge dx^{i_k}) = (dx^{j_1} \wedge \cdots \wedge dx^{j_{n-k}})$$

and extending it linearly, where $(i_1, \dots, i_k, j_1, \dots, j_{n-k})$ is an even permutation of $(1, 2, \dots, n)$ so that $dx^{i_1} \wedge \cdots \wedge dx^{i_k} \wedge dx^{j_1} \wedge \cdots \wedge dx^{j_{n-k}} = d\text{vol}$. Let $d^* : \Lambda^k \to \Lambda^{k-1}$ be the adjoint of the exterior differential operator $d : \Lambda^{k-1} \to \Lambda^k$ with respect to the *Hodge inner product*:

$$\langle \alpha, \beta \rangle := \int_\Omega \alpha \wedge *\beta, \quad \text{where} \quad \alpha, \beta \in L^2(\Omega; \Lambda^k). \tag{4.1}$$

More precisely, it is defined by $\langle d\alpha, \beta \rangle = \langle \alpha, d^*\beta \rangle$ for smooth forms $\alpha \in \Lambda^{k-1}(\Omega)$, $\beta \in \Lambda^k(\Omega)$, one of which with compact support in Ω. From the Stokes theorem, it follows that $d^* = (-1)^{nk+n+1} * d *$.

Let $u = (u^1, \dots, u^n) \in W^{1,2}(\Omega; \mathbb{R}^n)$. For the sake of simplicity, we will use the same notation u for the corresponding 1-form $\sum_{i=1}^n u^i(x)dx^i$. In this context, we denote its exterior differential du by

$$du := \sum_{i<j} (D_i u^j - D_j u^i) dx^i \wedge dx^j.$$

A celebrated result by De Giorgi [1] states that weak solutions to linear elliptic equations with L^∞ coefficients are Hölder continuous. In contrast to this, as it is well known, weak solutions of linear elliptic systems with L^∞ coefficients may have

singularities. For example, De Giorgi [2] constructed a weak solution to an elliptic system with L^∞ coefficients which belongs to $W^{1,2}(B_1(0);\mathbb{R}^n)$, $n \geq 3$, but is not bounded.

Related to those results, Giaquinta and Hong [4] raised an interesting question: Are weak solutions of the following system locally Hölder continuous?

$$\left.\begin{array}{c} d^*[a(x)du] = 0 \\ d^*u = 0 \end{array}\right\} \quad \text{in} \quad \Omega. \tag{4.2}$$

Here $a(x) \in L^\infty(\Omega)$ is assumed to be bounded by two positive numbers m, M. More generally, consider the following inhomogeneous system:

$$\left.\begin{array}{c} d^*[a(x)du] = f + d^*g \\ d^*u = 0 \end{array}\right\} \quad \text{in} \quad \Omega, \tag{4.3}$$

where $f \in \mathcal{H}^p_{loc}(\Omega;\mathbb{R}^n)$ and $g \in L^q_{loc}(\Omega;\Lambda^2) \cong L^q_{loc}(\Omega;\mathbb{R}^{n(n-1)/2})$.

When $n = 3$ the above system (4.3) is identical to the system (2.1), and Theorem 2.1 states the answer to their question is positive when $n = 3$. However, our method used in the proof of Theorem 2.1 cannot be applied to the case when $n \geq 4$, and we don't know the answer in that case.

Let us briefly mention why the case $n = 3$ is special. In the proof of Theorem 2.1, we made use of the fact that de Rham cohomology of a ball $B \in \mathbb{R}^n$ is trivial in the sense that if $\alpha \in \Lambda^2(B)$ satisfies $d^*\alpha = 0$, then there exists a $\beta \in \Lambda^{n-3}(B)$ such that $d\beta = *\alpha$. In the case when $n = 3$, β is a scalar function so that we may apply the well-known result of De Giorgi [1] to get the $C^{0,\alpha}$ estimate.

The aim of this section is to compile known results from general theory of elliptic systems which can be applied to the system (4.3). We have the following identity similar to (2.4) (see, e.g., [8, p. 33]):

$$-\Delta\alpha = d^*(d\alpha) + d(d^*\alpha) \quad \forall\alpha \in \Lambda^1(\Omega). \tag{4.4}$$

Hence, if $a(x)$ is continuous, then the perturbation method used in Lemma 2.9 can be applied here without any change. Also, if the ratio M/m is sufficiently close to 1, then it can be shown that weak solutions u of the system (4.3) satisfy $u \in W^{1,p}_{loc}$ for some $p > n$. Hölder continuity of u will then follow from Sobolev imbedding.

We again emphasize that most of results in this section can be inferred from the general theories of elliptic systems, so we will provide proofs only when the situation is not quite obvious.

PROPOSITION 4.1. *Let $u \in W^{1,2}_{loc}(\Omega;\mathbb{R}^n)$ be a weak solution of (4.3). Suppose $f \in \mathcal{H}^p_{loc}(\Omega;\mathbb{R}^n)$, $p > n/2$, and $g \in L^q_{loc}(\Omega;\Lambda^2)$, $q > n$. If $a(x)$ is continuous, then u is locally Hölder continuous with exponent $\alpha = \alpha(n, m, M, p, q) > 0$.*

Proof. See Theorem 3.1 in [3] and the following remark on page 87. □

PROPOSITION 4.2. *Let $u \in W^{1,2}_{loc}(\Omega;\mathbb{R}^n)$ be a weak solution of (4.3). Suppose $f \in \mathcal{H}^{q/2}_{loc}(\Omega;\mathbb{R}^n)$ and $g \in L^q_{loc}(\Omega;\Lambda^2)$, $q > n$. Then there exists a number $\epsilon_0 > 1$ such that if $M/m < \epsilon_0$, then $\nabla u \in L^p_{loc}(\Omega;\mathbb{R}^{n^2})$ for some $p > n$. In particular, u is locally Hölder continuous in Ω.*

Proof. The proof relies on the L^p theory for the Laplace operator and a perturbation argument (see, e.g., [6] and Theorem 2.5 (page 154) in [3]). □

PROPOSITION 4.3. *Let $u \in W^{1,2}_{loc}(\Omega;\mathbb{R}^n)$ be a weak solution of (4.3). Assume $f \in \mathcal{H}^p_{loc}(\Omega;\mathbb{R}^n)$, $p > n$, and $g \in C^{0,\beta}_{loc}(\Omega;\Lambda^2)$, $\beta > 0$. If $a(x)$ is $C^{0,\mu}$-continuous, then ∇u is locally Hölder continuous with exponent $\alpha = \min(\mu, 1 - n/p, \beta)$.*

Proof. See Theorem 3.2 (page 88) in [3] and also Lemma 2.9 in section 2. □

LEMMA 4.4 (Caccioppoli inequality). *Let $u \in W^{1,2}_{loc}(\Omega;\mathbb{R}^n)$ be a weak solution of the system* (4.3) *with $f \in \mathcal{H}^{2n/(n+2)}_{loc}(\Omega;\mathbb{R}^n)$ and $g \in L^2_{loc}(\Omega;\Lambda^2)$. Let $B_R := B_R(x_0) \Subset \Omega$. Then, for any $\lambda \in \mathbb{R}^n$,*

$$\int_{B_{R/2}} |\nabla u|^2 \le C\left(\frac{1}{R^2}\int_{B_R}|u-\lambda|^2 + \|f\|^2_{L^{2n/(n+2)}(B_R)} + \|g\|^2_{L^2(B_R)}\right), \tag{4.5}$$

where $C = C(n,m,M)$.

Sketch of proof. As in Lemma 2.4, there exists $h \in W^{1,2}_0(B_R;\Lambda^2)$ such that

$$f = d^*h \quad \text{in} \quad B_R; \qquad \|h\|_{L^2(B_R)} \le C\,\|f\|_{L^{2n/(n+2)}(B_R)}. \tag{4.6}$$

Let $\eta \in C^\infty_0(B_R;\mathbb{R})$ be a cut-off function such that $0 \le \eta \le 1$, $\eta \equiv 1$, in $B_{R/2}$ and $|\nabla\eta| \le 4/R$. By choosing $(u-\lambda)\eta^2$ as a test function it is easy to see that

$$\int_{B_R}\eta^2\,|du|^2 \le C\left(\frac{1}{R^2}\int_{B_R}|u-\lambda|^2 + \int_{B_R}|h|^2 + \int_{B_R}|g|^2\right). \tag{4.7}$$

Since $d^*u = 0$ in Ω, (4.4) implies

$$\begin{aligned}\big\langle -\Delta u, \eta^2(u-\lambda)\big\rangle &= \big\langle du, d\left(\eta^2(u-\lambda)\right)\big\rangle \\ &= \langle du, 2\eta\,d\eta\wedge(u-\lambda)\rangle + \big\langle du, \eta^2 du\big\rangle.\end{aligned}$$

On the other hand, integration by parts yields

$$\begin{aligned}\big\langle -\Delta u, \eta^2(u-\lambda)\big\rangle &= \int_\Omega \nabla u\cdot\nabla\left(\eta^2(u-\lambda)\right) \\ &= \int_\Omega 2\eta\,D_j u^i D_j\eta(u^i-\lambda^i) + \int_\Omega \eta^2\,|\nabla u|^2.\end{aligned}$$

Therefore

$$\int_{B_R}\eta^2\,|\nabla u|^2 \le C\left(\frac{1}{R^2}\int_{B_R}|u-\lambda|^2 + \int_{B_R}\eta^2\,|du|^2\right). \tag{4.8}$$

Combining (4.6), (4.7), and (4.8) we obtain (4.5). □

LEMMA 4.5 (L^p estimates). *Suppose $f \in \mathcal{H}^q_{loc}(\Omega;\mathbb{R}^n)$, $q > 2n/(n+2)$, and $g \in L^r_{loc}(\Omega;\Lambda^2)$, $r > 2$. Let $u \in W^{1,2}_{loc}(\Omega)$ be a weak solution of the system* (4.3). *Then $\nabla u \in L^p_{loc}(\Omega;\mathbb{R}^{n^2})$ for some $p > 2$. More precisely, let $B := B_R(x_0) \Subset \Omega$; then*

$$\|\nabla u\|_{L^p(B_{R/2})} \le C\left(\|\nabla u\|_{L^2(B)} + \|f\|_{L^{np/(n+p)}(B)} + \|g\|_{L^p(B)}\right). \tag{4.9}$$

Sketch of proof. Let h be as in (4.6). Setting $\lambda = (u)_R$ and then using Sobolev–Poincaré inequality, we obtain from (4.7) and (4.8)

$$⨍_{B_{R/2}}|\nabla u|^2 \le C\left[\left(⨍_B|\nabla u|^s\right)^{2/s} + ⨍_B|h|^2 + ⨍_B|g|^2\right], \quad s = \frac{2n}{n+2}. \tag{4.10}$$

It is so-called reverse Hölder inequality. It is well known that higher integrability of ∇u follows from (4.10) (see, e.g., Proposition 1.1 (page 122) of [3]). Also, as mentioned

in Proposition 4.2, (4.9) can be derived by a perturbation argument based on the L^p theory of the Laplace operator. □

With the preceding lemmas at hand, let us consider the quasi-linear system

$$\left.\begin{array}{c} d^*[\sigma(x,u)du] = f + d^*g \\ d^*u = 0 \end{array}\right\} \quad \text{in} \quad \Omega, \tag{4.11}$$

where $f \in \mathcal{H}^p_{loc}(\Omega)$ and $g \in L^q_{loc}(\Omega;\Lambda^2)$.

By using general theory of elliptic systems, it is again more or less straightforward to show partial $C^{0,\alpha}$ (or $C^{1,\alpha}$) regularity for weak solutions of the system (4.11) under appropriate continuity assumptions on σ. We denote k-dimensional Hausdorff measure of $\Sigma \subset \mathbb{R}^n$ by $H^k(\Sigma)$.

PROPOSITION 4.6 ($C^{0,\alpha}$-partial regularity). *Suppose $f \in \mathcal{H}^{q/2}_{loc}(\Omega)$ and $g \in L^q_{loc}(\Omega)$, for some $q > n$, and let $u \in W^{1,2}_{loc}(\Omega)$ be a weak solution of the system* (4.11). *Assume that σ is continuous. Then there exists an open set $\Omega_0 \subset \Omega$ such that u is locally Hölder continuous with exponent $1 - n/q$ in Ω_0. Moreover, $H^{n-s}(\Omega \setminus \Omega_0) = 0$ for some $s > 2$.*

Proof. See Theorem 1.1 (page 166) in [3]. □

PROPOSITION 4.7 ($C^{1,\alpha}$-partial regularity). *Suppose σ is locally $C^{0,\alpha}$-continuous for some $\alpha \in (0,1)$. Let $u \in W^{1,2}_{loc}(\Omega)$ be a weak solution of the system* (4.11) *and let $f \in \mathcal{H}^p_{loc}(\Omega;\mathbb{R}^n)$, $p = n/(1-\alpha)$, $g \in C^{0,\alpha}_{loc}(\Omega;\mathbb{R}^n)$. Then there exists an open set $\Omega_0 \subset \Omega$ such that $u \in C^{1,\alpha}_{loc}(\Omega_0)$ and $H^{n-s}(\Omega \setminus \Omega_0) = 0$ for some $s > 2$.*

Proof. The proof follows from Proposition 4.6 and Lemma 2.9. □

Acknowledgments. We thank Professor Min-Chun Hong and Professor Mariano Giaquinta for valuable communications. We also thank Professor Mikhail Safonov and Professor Vladimír Šverák for helpful discussions.

REFERENCES

[1] E. De Giorgi, *Sulla differenziabilità e l'analiticità delle estremali degli integrali multipli regolari*, Mem. Accad. Sci. Torino. Cl. Sci. Fis. Mat. Nat. (3), 3 (1957), pp. 25–43.

[2] E. De Giorgi, *Un esempio di estremali discontinue per un problema variazionale di tipo ellittico*, Boll. Un. Mat. Ital. (4), 1 (1968), pp. 135–137.

[3] M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, Ann. Math. Stud. 105, Princeton University Press, Princeton, NJ, 1983.

[4] M. Giaquinta and M.-C. Hong, *Partial Regularity of Minimizers of a Functional Involving Forms and Maps*, preprint.

[5] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order,* 2nd ed., Springer-Verlag, Berlin, 1983.

[6] N. G. Meyers, *An L^p-estimate for the gradient of solutions of second order elliptic divergence equations*, Ann. Scuola Norm. Sup. Pisa (3), 17 (1963), pp. 189–206.

[7] L. C. Piccinini and S. Spagnolo, *On the Hölder continuity of solutions of second order elliptic equations in two variables*, Ann. Scuola Norm. Sup. Pisa (3), 26 (1972), pp. 391–402.

[8] S. Rosenberg, *The Laplacian on a Riemannian Manifold*, Cambridge University Press, Cambridge, UK, 1997.

[9] H.-M. Yin, *Regularity of solutions to Maxwell's system in quasi-stationary electromagnetic fields and applications*, Comm. Partial Differential Equations, 22 (1997), pp. 1029–1053.

[10] H.-M. Yin, *On Maxwell's equations in an electromagnetic field with the temperature effect*, SIAM J. Math. Anal., 29 (1998), pp. 637–651.

[11] H.-M. Yin, *Optimal regularity of solution to a degenerate elliptic system arising in electromagnetic fields*, Commun. Pure Appl. Anal., 1 (2002), pp. 127–134.

SIAM J. MATH. ANAL.
Vol. 34, No. 1, pp. 101–120

FUNCTIONS AND DOMAINS HAVING MINIMAL RESISTANCE UNDER A SINGLE-IMPACT ASSUMPTION*

M. COMTE† AND T. LACHAND-ROBERT†

Abstract. We are looking for the domains $\Omega \subset \mathbb{R}^2$ tiling the plane and for functions $u : \Omega \to \mathbb{R}$ satisfying the simple impact assumption introduced by G. Buttazzo, V. Ferone, and B. Kawohl [*Math. Nach.*, 173 (1993), pp. 71–89.] about Newton's problem of the body of minimal resistance, which minimizes functionals $F(u;\Omega) = \frac{1}{|\Omega|}\int_\Omega f(|\nabla u|)$, with f decreasing.

We prove that only some convex polygons are minimizers, and we give explicitly the corresponding functions u. In the case of the Newton's functional $f(t) = 1/(1+t^2)$, all optimal domains are squares or regular hexagons.

Key words. body of minimal resistance, Newton's problem, single-impact, calculus of variations

AMS subject classifications. 49K30, 65K10

PII. S0036141001388841

1. Introduction.

Newton's problem of the body of minimal resistance, first stated in [6], has been widely studied, and new interest has recently been raised, particularly since the minimizer has been proven to not have the symmetry of the domain. This problem can be mathematically stated as follows: minimize

$$G(u) = \int_\Omega \frac{dx}{1 + |\nabla u(x)|^2}$$

in an appropriate class of functions $u : \Omega \to \mathbb{R}$. Here $\Omega \subset \mathbb{R}^2$ is a bounded domain, and the functional expresses the resistance of the body in $\{(x,z) \in \Omega \times \mathbb{R};\ z \le u(x)\}$ to a uniform stream of particles coming downward in the vertical direction. It is assumed that each particle's shock on the body is perfectly elastic and that each particle hits the body at most once. The elastic assumption leads on the given value of the functional [2]. Since the infimum of G is clearly zero, if we can consider very long and thin bodies, it is necessary to restrict the class of admissible functions u by fixing the surface area, or fixing the maximal height as a number $M > 0$, a given parameter. This last requirement was first proposed by Newton and is the most frequently considered one.

The single impact assumption is classically enforced by considering only concave functions u. On the other hand, it was shown in [2] that this requirement is not necessary. More precisely, it is sufficient to consider the unknown function u in the class of maps $u : \Omega \to [-M, 0]$ satisfying the geometrical condition

$$(1)\quad \forall x \in \mathrm{dom}(\nabla u),\ \forall \tau > 0,\ \text{such that } x - \tau\nabla u(x) \in \overline{\Omega},$$
$$\frac{u(x - \tau\nabla u(x)) - u(x)}{\tau} \le \frac{1}{2}\left(1 - |\nabla u(x)|^2\right).$$

Here $\mathrm{dom}(\nabla u)$ is merely the dense subset of Ω where u is differentiable; we will define it more precisely later.

*Received by the editors April 27, 2001; accepted for publication (in revised form) February 1, 2002; published electronically August 15, 2002.
http://www.siam.org/journals/sima/34-1/38884.html
†Université Pierre et Marie Curie, Laboratoire d'Analyse Numérique, 75252 Paris Cedex 05, France (comte@ann.jussieu.fr, lachand@ccr.jussieu.fr, www.ann.jussieu.fr).

This constraint has been studied in [4] and [3], where it is shown that a radial minimizer exists (but there is generally no uniqueness), but in the general class the minimum is not attained. This comes from the presence of a boundary, which induces special effects, and, in particular, allows oscillations near it. This leads us to the idea that without a boundary, the problem would be more "stable," and a global minimizer could exist with no radial symmetry assumption.

In this paper we investigate this problem. In order to make sense, we consider an infinite body with periodicity; that is, let Ω be a domain *tiling the plane* (meaning that there exists a finitely generated subgroup $\mathcal{G}_\Omega \subset \mathcal{O}_2$ such that $\bigcup_{g\in\mathcal{G}_\Omega} g(\overline{\Omega}) = \mathbb{R}^2$, and $g_1(\Omega) \cap g_2(\Omega) = \emptyset$ if $g_1 \neq g_2$); let $\bar{u} : \mathbb{R}^2 \to \mathbb{R}$ be a function with the same periodicity ($\bar{u} \circ g = \bar{u}$ for all $g \in \mathcal{G}_\Omega$); and let u be the restriction of $\bar{u}$ to Ω. We are looking to the body $\{(x, z) \in \mathbb{R}^3;\ z \leq \bar{u}(x)\}$ minimizing the mean value of G, that is, we minimize

$$F(u;\Omega) := \frac{1}{|\Omega|}\int_\Omega \frac{dx}{1+|\nabla u(x)|^2} \tag{2}$$

with respect to all domains Ω tiling the plane and to all functions $u : \Omega \to [-M, 0]$ having periodicity Ω.

Note that Ω is not well defined in general if only $\bar{u}$ is given. In order to fix the notations, we choose Ω such that $u(x) = 0$ for all $x \in \partial\Omega$ and $u < 0$ in Ω.

The natural topology associated with (1) is $W^{1,\infty}(\Omega)$. However it has been shown in [3] that the set of admissible functions must be restricted to a strict subset of $W^{1,\infty}(\Omega)$. Moreover, the minimizing functions must be *regular* in the following sense: these are continuous, C^2 by parts, functions; that is, they are obtained by a finite number of min or max operations on C^2 functions. In this paper, we do not enter these technicalities, and we restrain ourselves to this smaller class of regular functions. Since $u = 0$ on $\partial\Omega$, this implies also some regularity on Ω itself.

Our main result reads as follows.

THEOREM 1. *Among all regular functions and regular domains tiling the plane, the minimum of F is attained in only two cases, up to a similitude (with the same minimal value):*

1. *Ω is a square, $\Omega = (-a, a) \times (-a, a)$ with $a \leq 4M/3$, and u is the function*

$$u(x_1, x_2) := \max[\phi_a(|x_1|), \phi_a(|x_2|)], \tag{3}$$

   where $\phi_a(x) := \frac{(x+a)^2}{4a} - a$.
2. *Ω is a regular convex hexagon with diameter $4a/\sqrt{3}$, with center $O = (0, 0)$, and two vertices $A = (a, a/\sqrt{3})$, $B = (a, -a/\sqrt{3})$; then u is the function invariant by rotation of $\pi/3$ whose restriction to the triangle OAB is $\phi_a(x_1)$.*

In both cases, the optimal value for F is given by

$$F_{\text{opt}} := \pi + 12\ln 2 - 4\ln 5 - 4\arctan 2 \simeq 0.5930123.$$

Hence the resistance of the infinite tiling is less than 60% of the resistance of the plane (which has maximal resistance).

The proof of this theorem constitutes the rest of the paper. It relies on the following properties: first of all, any minimizing domain Ω is convex. This is proved in Theorem 2 hereafter, even for a more general functional $\frac{1}{|\Omega|}\int_\Omega f(|\nabla u|)$, with f decreasing. Section 2 states this theorem and gives the proof.

Since Ω should also tile the plane, it is a polygon. In section 3, we restrict ourselves to the functional of (2), give additional properties of optimal polygons, and characterize the corresponding function u. This allows us to restrict ourselves to a small set of admissible polygons. We end by using some explicit computations to distinguish the minimizers.

2. Convexity of the optimal domains. The first fact of interest is that any minimizer u of (2) saturates everywhere in $\mathrm{dom}(\nabla u)$; that is, for all $x \in \mathrm{dom}(\nabla u)$ there exists $\tau_x > 0$ such that

$$\frac{u(x - \tau_x \nabla u(x)) - u(x)}{\tau_x} = \frac{1}{2}\left(1 - |\nabla u(x)|^2\right), \tag{4}$$

and for all $\tau \in (0, \tau_x)$ inequality (1) is strict. This has been proved in [4], using a small variation with a fast oscillating function. Since the proof is quite lengthy and technical and works similarly here, we don't repeat it.

Note that (4) implies that

$$|\nabla u| \le 1 \tag{5}$$

and $x - \tau_x \nabla u \in \overline{\Omega}$ if $x \in \Omega$. Indeed, x is an interior point of the bounded set Ω; hence there exists $\tau_1 > 0$ such that $x - \tau_1 \nabla u(x)$ belongs to $\partial\Omega$. Since we assume $u = 0$ on $\partial\Omega$, we get $u(x - \tau_1 \nabla u(x)) = 0 \ge u(x)$. Putting this inequality into (1) yields (5) as claimed. Now if $\tau > \tau_1$,

$$\begin{aligned} u(x) + \frac{\tau}{2}\left(1 - |\nabla u(x)|^2\right) &> u(x) + \frac{\tau_1}{2}\left(1 - |\nabla u(x)|^2\right) \\ &\ge u(x - \tau_1 \nabla u(x)) = 0. \end{aligned}$$

This implies $\tau_x \le \tau_1$ since u takes values in $[-M, 0]$; hence $x - \tau_x \nabla u(x) \in \overline{\Omega}$.

Also this implies that

$$|\nabla u(x)| = 1 \qquad \forall x \in \partial\Omega \cap \mathrm{dom}(\nabla u). \tag{6}$$

Indeed, for $x \in \partial\Omega$, $u(x) = 0$, so the left-hand side of (4) is nonpositive. Hence $|\nabla u(x)| \ge 1$ and (6) follows since $|\nabla u| \le 1$.

Remark 2.A. First we have that if $x \in \partial\Omega \cap \mathrm{dom}(\nabla u)$, then $u(x) = 0$ and $u(x - \tau_x \nabla u(x)) = 0$, and from the definition of Ω, $x - \tau_x \nabla u(x) \in \partial\Omega$.

We also obtain that (5) implies

$$1 \ge F(u; \Omega) \ge \frac{1}{2} \tag{7}$$

for all admissible $(u; \Omega)$.

The main theorem relies primarily on the following result, which can be stated for a more general functional.

THEOREM 2. *Let $(u; \Omega)$ be a local minimizer pair for a functional in the form $\Upsilon(u; \Omega) := \frac{1}{|\Omega|}\int_\Omega f(|\nabla u|)$, where $f \in C^1(\mathbb{R}_+)$ is decreasing. Then Ω is convex.*

Proof. Assume that Ω is not convex. Then there exists $x_0 \in \partial\Omega$, a straight line Δ containing x_0, and a neighborhood V of x_0 such that, if P_1, P_2 are the two open half-planes limited by Δ,

$$V \cap P_1 \subset \Omega \quad \text{and} \quad V \cap P_2 \cap \Omega \ne \emptyset. \tag{8}$$

Choosing the appropriate center of coordinates, we can assume that $x_0 = 0$; also we can assume that 0 is an exposed point on the boundary of $\mathbb{R}^2 \setminus \Omega$, that $0 \in \mathrm{dom}(\nabla u)$, and that

$$\Delta \cap V \cap \partial\Omega = \{0\}. \tag{9}$$

We recall that $u(0) = 0$ and, since $|\nabla u(0)| = 1$, we can assume $\nabla u(0) = -e_1$ again by choosing the appropriate coordinate system, which is the unit outward normal vector to Ω at 0. Let τ_0 be the value of τ_x corresponding to $x = 0$.

Step 1. Let us first assume that u is not differentiable near 0. More precisely, V can be divided in two zones V_1, V_2, and there exists two functions u_1, u_2 of class $C^2(V)$ such that $u = u_i$ on V_i, $i = 1, 2$. Since u is continuous, $u_1 = u_2$ on the common boundary of V_1, V_2. We have $(u_1 - u_2)(0) = 0$, but the assumption that u is not differentiable at 0 implies in particular that $\nabla u_1(0) \neq \nabla u_2(0)$. Since $u_1 - u_2$ is C^2, the line $u_1 = u_2$ is a differentiable curve containing 0. Let us choose a coordinate system such that e_1 is tangent to this curve, pointing inside Ω at 0.

Since $|\nabla u| = 1$ on $\partial\Omega$, we can write that

$$u_1(x) = -\cos\alpha\ x_1 - \sin\alpha\ x_2 + o(|x_1| + |x_2|)$$

in V. Also

$$u_2(x) = -\cos\alpha\ x_1 + \sin\alpha\ x_2 + o(|x_1| + |x_2|)$$

in V, taking into account that $u_1 = u_2$ on the line $x_2 = 0$. Moreover, $\cos\alpha > 0$ since e_1 points inside Ω. Hence, if the direction of the axis x_2 is chosen appropriately, we have $u(x) = -\cos\alpha\ x_1 - \sin\alpha\ |x_2| + o(|x_1| + |x_2|)$ near 0.

Let us now define $\theta_0(x) = -\cos\beta\, x_1 - \sin\beta\, |x_2|$, where β satisfies $\cos\beta \in (0, \cos\alpha)$ and $\cos(\alpha - \beta) > 0$. For any $\varepsilon > 0$, we define $\theta_\varepsilon(x) := (1 - \delta_\varepsilon)\theta_0(x_1 + \varepsilon, x_2)$, where $\delta_\varepsilon > 0$ satisfies $\lim_{\varepsilon\to 0} \delta_\varepsilon = 0$ and $\lim \frac{\varepsilon}{\delta_\varepsilon} = 0$.

We extend u outside Ω by zero, and define V_ε as the connected part of $\{x; \theta_\varepsilon(x) < u(x)\}$ containing 0. Note that V_ε is approximately equal to the set

$$T_\varepsilon := \Big\{ x \in \mathbb{R}^2; \\ \max(0, -\cos\alpha\ x_1 - \sin\alpha\ |x_2|) > -(1 - \delta_\varepsilon)(\cos\beta\ (x_1 + \varepsilon) + \sin\beta\, |x_2|) \Big\},$$

which is a triangle symmetric with respect to $\{x_2 = 0\}$. Note that the edges of T_ε have length of order ε, so that $|V_\varepsilon| \sim |T_\varepsilon| \sim c\varepsilon^2$ for some constant c. Also we can assume that there exists $C_1 > 0$ such that $V_\varepsilon \subset B(0, C_1\varepsilon)$ for ε small enough. In particular there exists $C_2 > 0$

$$\forall x \in V_\varepsilon, \qquad |\theta_\varepsilon(x)| < C_2\varepsilon. \tag{10}$$

We now define $\Omega_\varepsilon := \Omega \cup V_\varepsilon$, and

$$u_\varepsilon = \begin{cases} \theta_\varepsilon & \text{in } V_\varepsilon, \\ u & \text{in } \Omega \setminus V_\varepsilon. \end{cases} \tag{11}$$

Let us prove that u_ε satisfies the constraint (1) if ε is small enough and if δ_ε is chosen appropriately.

If x and y belong to $\Omega \setminus \overline{V}_\varepsilon$, then $u_\varepsilon \equiv u$ near x and y, so the constraint for u_ε follows from (1). If x and y belong to V_ε, then u_ε is affine near x and y; hence it satisfies (1).

If $x \in \Omega \setminus \overline{V}_\varepsilon$, $y \in V_\varepsilon$, then $u_\varepsilon(y) = \theta_\varepsilon(y) \le u(y)$, and $u_\varepsilon \equiv u$ near x, so the constraint follows from (1) again.

The last case to consider is $x \in V_\varepsilon$, $y \in \Omega \setminus V_\varepsilon$. We have $\nabla\theta_0(x) = -\cos\beta\ e_1 - \sin\beta\ \operatorname{sgn}(x_2)e_2$; hence

$$\frac{d}{dt}u(x - t\nabla\theta_0(x))\Big|_{t=0} = -\nabla u(x)\cdot\nabla\theta_0(x) = -\cos(\alpha-\beta) < 0$$

by assumption. Hence, reducing V if necessary, we can assume that for all $x \in V$ the map $t \mapsto u(x - t\nabla\theta_0(x))$ is decreasing for $t > 0$ satisfying $x - t\nabla\theta_0(x) \in V$.

Now if $x \in V_\varepsilon$, $y = x - \tau\nabla\theta_\varepsilon(x) = x - t\nabla\theta_0(x)$, where $t = \tau(1-\delta_\varepsilon)$, we distinguish two cases: if $\tau > C_2\varepsilon/\delta_\varepsilon(1-\delta_\varepsilon)$, we have, using (10), that

$$\theta_\varepsilon(x) + \tau(1 - |\nabla\theta_\varepsilon|^2) > -C_2\varepsilon + \tau\left(\delta_\varepsilon - \frac{1}{2}\delta_\varepsilon^2\right) > 0 \ge u_\varepsilon(y);$$

hence the constraint is satisfied.

In the other case, using $V_\varepsilon \subset B(0, C_1\varepsilon)$,

$$|y| \le |x| + |x-y| \le C_1\varepsilon + \tau(1-\delta_\varepsilon) = \varepsilon C_1 + C_2\frac{\varepsilon}{\delta_\varepsilon} \longrightarrow 0 \text{ as } \varepsilon \to 0.$$

Therefore, if ε is small enough, that will imply $y \in V$. We recall that this implies that the map $t \mapsto u(x - t\nabla\theta_0(x))$ is decreasing. Since $y \notin V_\varepsilon$, there exists $t' \in (0,t)$ such that $z := x - t'\nabla\theta_0(x)$ belongs to the boundary of V_ε; then

$$u_\varepsilon(y) = u(y) \le u(z) = \theta_\varepsilon(z) \le \theta_\varepsilon(x).$$

Hence u_ε satisfies the constraint since $\frac{1}{2}(1 - |\nabla u_\varepsilon(x)|^2) > 0$.

We now compute $\delta\Upsilon := \Upsilon(u_\varepsilon;\Omega_\varepsilon) - \Upsilon(u;\Omega)$. In the following, we note $W_\varepsilon := V_\varepsilon \cap \Omega$, and $X_\varepsilon := V_\varepsilon \setminus W_\varepsilon$. Then $|\Omega_\varepsilon| = |\Omega| + |X_\varepsilon|$. We have $|V_\varepsilon| \sim c\varepsilon^2$ as $\varepsilon \to 0$, as explained before, and also $|X_\varepsilon| \sim c_1\varepsilon^2$, $|W_\varepsilon| \sim (c - c_1)\varepsilon^2$ for some constant c_1.

We have

$$\begin{aligned}
\delta\Upsilon &= \frac{1}{|\Omega_\varepsilon|}\left(\int_{V_\varepsilon} f(|\nabla\theta_\varepsilon|) + \int_{\Omega\setminus V_\varepsilon} f(|\nabla u|)\right) - \frac{1}{|\Omega|}\int_\Omega f(|\nabla u|) \\
&= \frac{1}{|\Omega_\varepsilon|}\left[\int_{X_\varepsilon} f(|\nabla\theta_\varepsilon|) + \int_{W_\varepsilon} f(|\nabla\theta_\varepsilon|) - f(|\nabla u|)\right] + \left[\frac{1}{|\Omega_\varepsilon|} - \frac{1}{|\Omega|}\right]\int_\Omega f(|\nabla u|) \\
&= \frac{|X_\varepsilon|}{|\Omega_\varepsilon|}\left[\frac{1}{|X_\varepsilon|}\int_{X_\varepsilon} f(|\nabla\theta_\varepsilon|) - \frac{1}{|\Omega|}\int_\Omega f(|\nabla u|) + R_\varepsilon\right] \\
&= \frac{|X_\varepsilon|}{|\Omega_\varepsilon|}\left[\Upsilon(\theta_\varepsilon; X_\varepsilon) - \Upsilon(u;\Omega) + R_\varepsilon\right],
\end{aligned}$$

where

$$R_\varepsilon := \frac{1}{|X_\varepsilon|}\int_{W_\varepsilon} f(|\nabla\theta_\varepsilon|) - f(|\nabla u|).$$

We deduce that $\lim_{\varepsilon\to 0} R_\varepsilon = 0$. Indeed, $|W_\varepsilon| / |X_\varepsilon|$ is bounded, and, since $\nabla\theta_\varepsilon = -(1-\delta_\varepsilon)e_1$ and $\nabla u = -e_1 + o(1)$ on V_ε, the result follows directly.

Going back to $\delta\Upsilon$, we note that

$$\lim_{\varepsilon\to 0} \Upsilon(\theta_\varepsilon; X_\varepsilon) = f(1)$$

since $|\nabla\theta_\varepsilon| = 1 + o(1)$ in V_ε. We conclude that there exists a constant $C > 0$ such that

$$\delta\Upsilon = C[f(1) - \Upsilon(u;\Omega)]\ \varepsilon^2 + o(\varepsilon^2).$$

Since u is a minimizer, $\delta\Upsilon \geq 0$; hence $\Upsilon(u;\Omega) \leq f(1)$. On the other hand, $|\nabla u| \leq 1$ in Ω and f is decreasing, so $f(|\nabla u|) \geq f(1)$. We deduce that $|\nabla u| = 1$ almost everywhere in Ω. This implies $u \geq 0$ in Ω from (1) and the fact that $u = 0$ on $\partial\Omega$ (consider a $\tau > 0$ such that $x - \tau\nabla u(x) \in \partial\Omega$). We already know that $u \leq 0$, so $u \equiv 0$, a contradiction.

This ends the proof of the theorem in the case where u is not differentiable at 0.

Step 2. Let us now assume that u is C^2 near 0. Then we can write it in the form

$$u(x_1, x_2) = -x_1 + \frac{1}{2}\alpha_1 x_1^2 + \frac{1}{2}\alpha_2 x_2^2 + o(x_1^2 + x_2^2). \tag{12}$$

We have taken into account that 0 is an exposed point, that we have (8), (9), and the fact that $u = 0$ on $\partial\Omega$ and $u < 0$ in Ω. Moreover, Δ is the tangent line to $\partial\Omega$ at 0, and from (8) we have $\alpha_2 < 0$. Since $|\nabla u(x_1, 0)| = |1 - \alpha_1 x_1 + o(x_1)| < 1$ for $x_1 > 0$, we also have $\alpha_1 > 0$.

Let θ_0 be the function defined by

$$\theta_0(x_1, x_2) = \begin{cases} -x_1 + \frac{1}{2}\beta x_1^2 & \text{if } x_1 < \frac{2}{\beta}, \\ 0 & \text{otherwise,} \end{cases}$$

where β is a number to be chosen later satisfying $\beta > \alpha_1$. As this is the negative part of a parabola whose focus point is at $(\frac{1}{\beta}, 0)$, θ_0 satisfies the constraint for every $x \in \mathbb{R}^2$ with $x_1 \geq 0$ and for every $\tau > 0$.

For any $\varepsilon > 0$, let $\theta_\varepsilon(x_1, x_2) := \theta_0(x_1 + \varepsilon^2, x_2)$. We claim that there exists $\varepsilon_0 > 0$, and, for any $\varepsilon \in (0, \varepsilon_0)$, a neighborhood $V_\varepsilon \subset V$ of 0 such that

$$\limsup_{\varepsilon\to 0} \frac{\operatorname{diam} V_\varepsilon}{\varepsilon} < \infty, \quad u > \theta_\varepsilon \text{ in } V_\varepsilon \cap \Omega, \quad \text{and} \quad u = \theta_\varepsilon \text{ on } \partial V_\varepsilon \cap \overline{\Omega}. \tag{13}$$

Let us extend u by the constant value 0 outside Ω and consider V_ε the connected component of $\{u > \theta_\varepsilon\}$ containing 0. Notice first that

$$V_\varepsilon \cap (\mathbb{R}^2 \setminus \Omega) \subset \{x_1 > -\varepsilon^2\}$$

since u was extended by zero.

We now consider $W_\varepsilon := V_\varepsilon \cap \overline{\Omega}$. Using the Taylor expansion of u, we have

$$\begin{aligned} u(x_1, x_2) - \theta_\varepsilon(x_1, x_2) = \varepsilon^2(1 - \beta x_1) - \frac{1}{2}(\beta - \alpha_1)x_1^2 \\ + \frac{1}{2}\alpha_2 x_2^2 - \frac{1}{2}\beta\varepsilon^4 + o(x_1^2 + x_2^2). \end{aligned} \tag{14}$$

Hence this vanishes for x_1, x_2 of order ε, and the set $u > \theta_\varepsilon$ is described by the equation

$$\varepsilon^2 + o(\varepsilon^2) > \frac{1}{2}(\beta - \alpha_1)\left(x_1 + \frac{\beta}{\beta - \alpha_1}\varepsilon^2\right)^2 - \frac{1}{2}\alpha_2 x_2^2.$$

Except for the $o(\varepsilon^2)$ term, this is the interior of an ellipse with center $(-\frac{\beta}{\beta-\alpha_1}\varepsilon^2, 0)$ and diameter of order ε. That proves that $\operatorname{diam} W_\varepsilon$ is of order ε. Moreover, V_ε is included in the intersection of the ellipse and $\{x_1 > -\varepsilon^2\}$. This ends the proof of (13). Notice also that (14) implies that $|u - \theta_\varepsilon| \le c\varepsilon^2$ on V_ε for some $c > 0$.

Step 3. We now define $\Omega_\varepsilon := \Omega \cup V_\varepsilon$, and

$$u_\varepsilon = \begin{cases} \theta_\varepsilon & \text{in } V_\varepsilon, \\ u & \text{in } \Omega \setminus V_\varepsilon. \end{cases} \tag{15}$$

We claim that

$$\Upsilon(u_\varepsilon; \Omega_\varepsilon) < \Upsilon(u; \Omega) \tag{16}$$

for ε small enough: this will contradict the minimality of $(u; \Omega)$ and ends the proof of the theorem.

Let us first prove that u_ε satisfies the constraints in Ω_ε. Let $x \in \Omega_\varepsilon \cap \operatorname{dom}(\nabla u_\varepsilon)$, and $\tau > 0$ such that $y := x - \tau \nabla u_\varepsilon(x) \in \overline{\Omega}_\varepsilon$. We have to prove that

$$\frac{u_\varepsilon(y) - u_\varepsilon(x)}{\tau} \le \frac{1}{2}(1 - |\nabla u_\varepsilon(x)|^2). \tag{17}$$

If x and y belong to $\Omega \setminus \overline{V}_\varepsilon$, then $u_\varepsilon \equiv u$ near x and y, and so (17) follows from (1). If x and y belong to V_ε, then $u_\varepsilon \equiv \theta_\varepsilon$ near x and y, and so (17) is obvious.

If $x \in \Omega \setminus \overline{V}_\varepsilon$, $y \in V_\varepsilon$, then $u_\varepsilon(y) = \theta_\varepsilon(y) \le u(y)$, and $u_\varepsilon \equiv u$ near x, and so (17) follows from (1) again.

The last case to consider is $x \in V_\varepsilon$, $y \in \Omega \setminus V_\varepsilon$.

Let us define, for all $x \in V$ and any $\varepsilon > 0$,

$$\Sigma_\varepsilon(x) := \left\{ y \in x - \mathbb{R}_+(\nabla u(x) + \sqrt{\varepsilon} B_1);\, \frac{u(y) - u(x)}{|y - x|} \ge \frac{1 - |\nabla u(x)|^2}{2\,|\nabla u(x)|} - \sqrt{\varepsilon} \right\},$$

where B_1 is the unit ball of $\mathbb{R}^2$. This is nonempty since it contains $x - \tau_x \nabla u(x)$. We now define $Q_\varepsilon := \Sigma_\varepsilon(V_\varepsilon)$. We claim that there exist $\varepsilon_0 > 0$ and $k > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$, $Q_\varepsilon \subset \{x_1 > k\}$. Indeed, if not, we can find sequences $(\varepsilon^n)_n$ going to 0, $(k^n)_n$ going to zero, and $x^n \in V_{\varepsilon^n}$, $y^n \in \Sigma(x_n)$ such that $y_1^n \le k_n$. This implies that x^n converges to 0; hence $\nabla u(x^n) \to -e_1$. Since $y^n \in x^n - \mathbb{R}_+(\nabla u(x^n) + \sqrt{\varepsilon^n} B_1)$, we deduce $(y_2^n - x_2^n)/(y_1^n - x_1^n) \to 0$ and $y_1^n \in [x_1^n, k^n]$ goes to zero for n large. So x^n and y^n are in V for n large, and using (12), we get

$$u(y^n) - u(x^n) = x_1^n - y_1^n + o(\varepsilon^n) \qquad \text{and} \qquad |y^n - x^n| = y_1^n - x_1^n + o(\varepsilon^n).$$

Since $y^n \in \Sigma_{\varepsilon^n}(x_n)$, we get

$$-1 + o(1) = \frac{u(y^n) - u(x^n)}{|y^n - x^n|} \ge \frac{1 - |\nabla u(x^n)|^2}{2\,|\nabla u(x^n)|} - \sqrt{\varepsilon^n} = o(1)$$

since $|\nabla u(x^n)| \to 1$. This is a contradiction.

The parameters β and k and linked together in our construction, but we can increase β without changing k. This is permitted by the monotonicity properties associated with these parameters: if β is increased (to a new value $\tilde{\beta} > \beta$, say), the corresponding function $\tilde{\theta}_0$ satisfies $\tilde{\theta}_0 \geq \theta_0$. This implies $\tilde{V}_\varepsilon \subset V_\varepsilon$ and $\tilde{Q}_\varepsilon \subset Q_\varepsilon \subset \{x_1 > k\}$.

In the following, we will assume that $\beta > 2/k$.

Let us now assume that we can find $x_\varepsilon \in V_\varepsilon$, $\tau_\varepsilon > 0$ such that the constraint is not satisfied for u_ε at x_ε, $y_\varepsilon := x_\varepsilon - \tau_\varepsilon \nabla u_\varepsilon(x_\varepsilon) = x_\varepsilon - \tau_\varepsilon \nabla \theta_\varepsilon(x_\varepsilon)$; that is,

$$\frac{u(y_\varepsilon) - \theta_\varepsilon(x_\varepsilon)}{\tau_\varepsilon} > \frac{1}{2}(1 - |\nabla \theta_\varepsilon(x_\varepsilon)|^2). \tag{18}$$

Note that $y_1^\varepsilon = x_1^\varepsilon + \tau_\varepsilon(1 - \beta(x_1^\varepsilon + \varepsilon^2))$, $y_2^\varepsilon = x_2^\varepsilon$.

Extracting a subsequence, we can assume that $\tau_\varepsilon \to \tau_0$. We have $\tau_0 > 0$ since if not, y_ε goes to 0 as x_ε does; hence y_ε belongs to V for ε small enough. Using a Taylor expansion of u near 0 in (18) and the particular form of the coordinates of y_ε, we get

$$\beta x_1^\varepsilon - 1 + o(|x^\varepsilon|) = \frac{1}{\tau_\varepsilon}(-y_1^\varepsilon + o(|y^\varepsilon|) - x_1^\varepsilon + o(|x^\varepsilon|)) > \beta x_1^\varepsilon + o(|x^\varepsilon|),$$

which is a contradiction. This proves $\tau_0 > 0$.

Using (18), we get

$$\begin{aligned}\frac{u(y_\varepsilon) - u(x_\varepsilon)}{\tau_\varepsilon} - \frac{1}{2}(1 - |\nabla u(x_\varepsilon)|^2) \\ > - \frac{u(x_\varepsilon) - \theta_\varepsilon(x_\varepsilon)}{\tau_\varepsilon} + \frac{1}{2}(|\nabla u(x_\varepsilon)|^2 - |\nabla \theta_\varepsilon(x_\varepsilon)|^2).\end{aligned}$$

Note that both terms in the right are $O(\varepsilon)$, so $y_\varepsilon \in \Sigma_\varepsilon(x_\varepsilon)$ for ε small enough. In particular, this implies $y_1^\varepsilon > k > 2/\beta$. Hence $\theta_\varepsilon(y^\varepsilon) \geq 0 \geq u(y^\varepsilon)$. From (18) we deduce that

$$\frac{\theta_\varepsilon(y_\varepsilon) - \theta_\varepsilon(x_\varepsilon)}{\tau_\varepsilon} > \frac{1}{2}(1 - |\nabla \theta_\varepsilon(x_\varepsilon)|^2).$$

Hence the pair $(x^\varepsilon, y^\varepsilon)$ violates the constraint for θ_ε, which is impossible from its definition. This ends the proof that u_ε satisfies the constraint.

We now compute $\delta\Upsilon := \Upsilon(u_\varepsilon; \Omega_\varepsilon) - \Upsilon(u; \Omega)$. In the following, we note $W_\varepsilon := V_\varepsilon \cap \Omega$ and $X_\varepsilon := V_\varepsilon \setminus W_\varepsilon$. Then $|\Omega_\varepsilon| = |\Omega| + |X_\varepsilon|$. From the definition of V_ε, the boundary of X_ε has two parts, one included in $\{x_1 = -\varepsilon^2\}$, the other one in $\partial\Omega$. Note that the boundary of Ω can be defined by the equation $u(x_1, x_2) = 0$; that is, using (12),

$$x_1 = \frac{1}{2}\alpha_2 x_2^2 + o(x_2^2).$$

We deduce that

$$|X_\varepsilon| \sim \int_{-\varepsilon^2}^0 2 \int_0^{\sqrt{2x_1/\alpha_2}} dx_2 \, dx_1 = \frac{4}{3}\sqrt{\frac{-2}{\alpha_2}} \, \varepsilon^3 + o(\varepsilon^3).$$

Note that $|V_\varepsilon| = O(\varepsilon^2)$ from (13).

We have

$$\begin{aligned}\delta\Upsilon &= \frac{1}{|\Omega_\varepsilon|}\left(\int_{V_\varepsilon} f(|\nabla\theta_\varepsilon|) + \int_{\Omega\setminus V_\varepsilon} f(|\nabla u|)\right) - \frac{1}{|\Omega|}\int_\Omega f(|\nabla u|)\\ &= \frac{1}{|\Omega_\varepsilon|}\left[\int_{X_\varepsilon} f(|\nabla\theta_\varepsilon|) + \int_{W_\varepsilon} f(|\nabla\theta_\varepsilon|) - f(|\nabla u|)\right] + \left[\frac{1}{|\Omega_\varepsilon|} - \frac{1}{|\Omega|}\right]\int_\Omega f(|\nabla u|)\\ &= \frac{|X_\varepsilon|}{|\Omega_\varepsilon|}\left[\frac{1}{|X_\varepsilon|}\int_{X_\varepsilon} f(|\nabla\theta_\varepsilon|) - \frac{1}{|\Omega|}\int_\Omega f(|\nabla u|) + R_\varepsilon\right]\\ &= \frac{|X_\varepsilon|}{|\Omega_\varepsilon|}\left[\Upsilon(\theta_\varepsilon; X_\varepsilon) - \Upsilon(u;\Omega) + R_\varepsilon\right],\end{aligned}$$

where

$$R_\varepsilon := \frac{1}{|X_\varepsilon|}\int_{W_\varepsilon} f(|\nabla\theta_\varepsilon|) - f(|\nabla u|).$$

We claim that $\lim_{\varepsilon\to 0} R_\varepsilon = 0$. Indeed, since $\nabla\theta_\varepsilon = -e_1 + o(1)$ and $\nabla u = -e_1 + o(1)$ on V_ε, we have $\lim R_\varepsilon = \lim \bar{R}_\varepsilon$, where

$$\bar{R}_\varepsilon := \frac{1}{|X_\varepsilon|}\int_{V_\varepsilon} f(|\nabla\theta_\varepsilon|) - f(|\nabla u|) = R_\varepsilon + \frac{1}{|X_\varepsilon|}\int_{X_\varepsilon} f(|\nabla\theta_\varepsilon|) - f(|\nabla u|).$$

Let us define $U(x) := \nabla u(x) + e_1$, $\Theta_\varepsilon(x) := \nabla\theta_\varepsilon(x) + e_1$. We note that there exists $c > 0$ such that $|U(x)| + |\Theta_\varepsilon(x)| < c\varepsilon$ for all $x \in V_\varepsilon$. Hence using a Taylor expansion of f near $-e_1$ in the form

$$f(|-e_1 + U|) = f(1) + \Phi\cdot U + O(|U|^2),$$

we get

$$\bar{R}_\varepsilon = \frac{1}{|X_\varepsilon|}\left[\int_{V_\varepsilon} \Phi\cdot(\Theta_\varepsilon(x) - U(x)) + O(\varepsilon^2)\,|V_\varepsilon|\right].$$

The integral vanishes since it is equal to $\int \Phi\cdot\nabla(\theta_\varepsilon - u)$, which is zero from the Green formula and the fact that $u = \theta_\varepsilon$ on ∂V_ε. Hence $\bar{R}_\varepsilon = O(\varepsilon)$ from our previous estimates and goes to zero as claimed.

Going back to $\delta\Upsilon$, we note that

$$\lim_{\varepsilon\to 0}\Upsilon(\theta_\varepsilon; X_\varepsilon) = f(1)$$

since $|\nabla\theta_\varepsilon| = 1 + o(1)$ in V_ε. We conclude that there exists a constant $C > 0$ such that

$$\delta\Upsilon = C[f(1) - \Upsilon(u;\Omega)]\,\varepsilon^3 + o(\varepsilon^3).$$

Using the same argument as used in the end of Step 1, we are led again to a contradiction.

This ends the proof of Theorem 2. □

3. Newton's functional. The remaining part of the proof of the main theorem relies more precisely on the exact value of the functional in (2). We divide it into a few lemmas.

LEMMA 3.1. *If (u, Ω) satisfies the constraint, with Ω convex and u saturating the constraints everywhere, then Ω is a convex polygon having at most six vertices and satisfying the following properties:*

1. *For every side $[A, B]$ of the polygon, and for any vertex C, the orthogonal projection of C on the line (AB) does not lie on the open segment (A, B).*
2. *All inner angles of the polygon are $\geq \pi/2$.*

Proof. We already know that Ω is convex from Theorem 2. Since Ω tiles the plane with a locally finite tiling, Ω is the interior of a convex polygon with at most six vertices (cf. [1]). Of course this gives us that u is C^2 on $\partial\Omega$ except at the vertices. We note that $A_0, \dots, A_{p-1}$ are the vertices of $\partial\Omega$, $p \leq 6$.

Step 4. Since u is constant ($= 0$) on $\partial\Omega$, $\nabla u(x)$ is orthogonal to $\partial\Omega$ at x; in particular, since $x - y(x) \in \mathbb{R}\nabla u(x)$, we have

$$\forall x \in \partial\Omega \cap \mathrm{dom}(\nabla u), \quad x - y(x) \text{ is orthogonal to } \partial\Omega \text{ at } x. \tag{19}$$

We recall that we assumed u to be C^2 by parts; the map $x \mapsto y(x) := x - \tau_x \nabla u(x)$ is defined almost everywhere. Let $x_0 \in \partial\Omega \setminus \{A_0, \dots, A_{p-1}\}$ and V be a neighborhood of x_0 such that u is $C^2(\overline{\Omega} \cap V)$. We assume by contradiction that $y(x_0) \notin \mathrm{dom}(\nabla u)$, that is, $y(x_0) \in \{A_0, \dots, A_{p-1}\}$, say $y(x_0) = A_0$, for instance.

We note that the map y is $C^1(V \cap \Omega)$. Indeed, since $u(x - \tau_x \nabla u(x)) = 0$, we have, using (4),

$$-2u(x) = \tau_x(1 - |\nabla u(x)|^2).$$

For any $x \in \Omega \cap \mathrm{dom}(\nabla u)$, we have $u(x) < 0$; since there exists $t \in \mathbb{R}$ such that $x - t\nabla u(x) \in \partial\Omega$, so that $u(x - t\nabla u(x)) = 0$, we get from (4) that $|\nabla u(x)| < 1$. Consequently τ is $C^1(V \cap \Omega)$, and so is y.

For $\varepsilon > 0$ small enough, the equation $u(x) = -\varepsilon$ defines a C^1-line $\mathcal{L}_\varepsilon$ in $V \cap \Omega$ with a parametric representation $s \in I_\varepsilon \subset \mathbb{R} \to \xi_\varepsilon(s)$. The map

$$y(\xi_\varepsilon(s)) = \xi_\varepsilon(s) - \tau(\xi_\varepsilon(s))\nabla u(\xi_\varepsilon(s))$$

takes values in $\partial\Omega$ near A_0. On the other hand, it is a C^1 map; hence it takes images in $[A_{p-1}, A_0]$ or $[A_0, A_1]$.

Moreover, $\mathcal{L}_\varepsilon$ is continuous with respect to ε and converges to $\mathcal{L}_0 := V \cap \partial\Omega \ni x_0$. From (19) and the convexity of Ω, $y(\mathcal{L}_0) \cap (A_{p-1}, A_0) \neq \emptyset$ and $y(\mathcal{L}_0) \cap (A_0, A_1) \neq \emptyset$. This is a contradiction, and this proves the first assertion of the lemma.

Step 5. We now prove that every interior angle at the vertices $A_0, \dots, A_{p-1}$ is $\geq \pi/2$. Since Ω is a convex polygon tiling the plane, that implies also that $p > 3$.

Indeed, let us assume by contradiction that the angle at, say, A_1 is $< \pi/2$. Let us consider $x \in (A_0, A_1)$ near A_1; from the previous step, $y(x)$ belongs to the orthogonal line to $[A_0, A_1]$ at x and to $\partial\Omega$, that is, $y(x) \in (A_1, A_2)$. As x varies continuously from A_1 to A_0, $y(x)$ still belongs to the same segment as proved before. Let B_0 be the limit of $y(x)$ as $x \to A_0$, so that $B_0 \in [A_1, A_2]$. Then the triangle $[A_0, A_1, B_0]$ is rectangular in A_0 and not flat. This implies

$$|A_1 - A_0| < |A_1 - B_0| \leq |A_1 - A_2|\,.$$

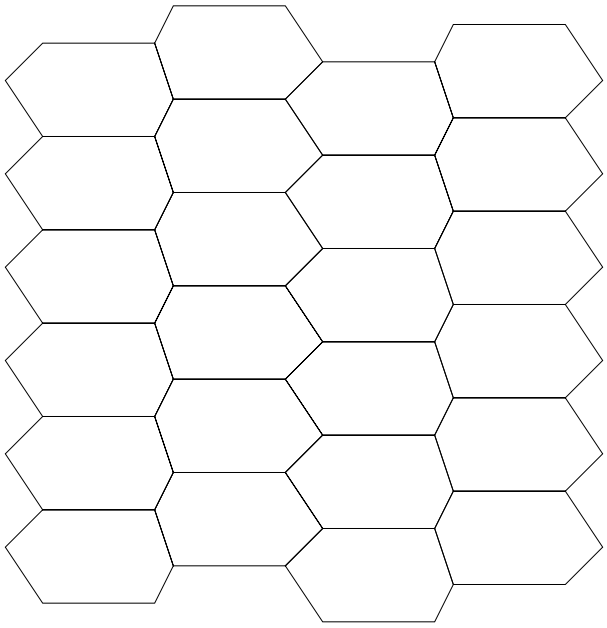


FIG. 1. *Tiling the plane with an irregular hexagon.*

On the other hand, A_2 obeys the same rule on the other side, and so $|A_1 - A_2| < |A_1 - A_0|$, which is a contradiction.

This ends the proof of the lemma. □

LEMMA 3.2. *The polygons obeying the properties of Lemma* 3.1 *and tiling the plane are*

1. *all the rectangles;*
2. *the regular hexagon;*
3. *all convex hexagons such that $A_1A_2A_4A_5$ is a rectangle, $\widehat{A}_0 \geq \pi/2$, $\widehat{A}_3 \geq \pi/2$, and $A_1A_2 > A_iA_{i+1}$ for $i = 2, 3, 5, 0$ (see Figure* 1*).*

Proof. Let us number the vertices such that A_1A_2 is one of the largest sides of the polygon. We shall write the indices modulo p, that is, $A_p = A_0$, etc. We call $\widehat{A}_i$ the inner angle at A_i.

Step 6. We claim that there exists $k > 1$ such that $A_1A_2A_kA_{k+1}$ is a rectangle.

Indeed, for any i, the orthogonal projection P_i of A_i on the line $\Delta = (A_1A_2)$ does not belong to (A_1, A_2), that is, belong to $\Delta_1 := \{A_1 + t(A_1 - A_2); t \in \mathbb{R}_+\}$ or $\Delta_2 := \{A_2 + t(A_2 - A_1); t \in \mathbb{R}_+\}$. Moreover, $P_0 \in \Delta_1$ since the inner angle at A_1 is $\geq \pi/2$; similarly, $P_3 \in \Delta_2$. Therefore, there exists k such that $P_k \in \Delta_2$ and $P_{k+1} \in \Delta_1$. Since the projection is 1-Lipschitz,

$$|A_k - A_{k+1}| \geq |P_k - P_{k+1}| \geq |A_1 - A_2|.$$

Since A_1A_2 is one of largest sides, there is equality here, and $P_k = A_2$, $P_{k+1} = A_1$.

Step 7. If $\widehat{A}_1 = \pi/2$, then the polygon is a rectangle. Indeed, $\widehat{A}_1 = \pi/2$ implies $A_{k+1} = A_0$, that is, $k = p-1$. If we suppose $\widehat{A}_2 > \pi/2$, then $k > 3$, and the projection of A_3 on (A_0, A_1) belongs to the interior of (A_0, A_1) from the convexity of the polygon and the fact that $A_0A_1A_2A_{p-1}$ is a rectangle. This contradicts Lemma 3.1.

Step 8. Let us now assume that the polygon is not a rectangle. From the previous

step, $\widehat{A}_i > \pi/2$ for $i = 1, 2, k, k+1$ since all four vertices of the rectangle $A_1A_2A_kA_{k+1}$ are equivalent. Consequently $k > 3$ and $k+1 < p$; since $p \leq 6$, this implies $p = 6$ and $k = 4$.

Let us first assume that $|A_2 - A_3| = |A_1 - A_2|$. Then using the same argument as in the first step, we get that $A_2A_3A_5A_0$ is a rectangle. Since this rectangle has a common diagonal (A_2A_5) with $A_1A_2A_4A_5$, both have the same center O. Therefore, A_i and A_{i+3} are symmetrical with respect to O for $i = 0, 1, 2$. The polygon is a regular hexagon.

Step 9. The last case to consider is $|A_1 - A_2| = |A_4 - A_5| > |A_i - A_{i+1}|$ for $i = 0, 2, 3, 5$.

Since Ω tiles the plane, there is a deplacement R such that A_1A_2 is a side of $R(\Omega)$; hence it must be $R(A_1A_2)$ or $R(A_4A_5)$. That implies that R is a translation or a rotation with angle π or a symmetry with respect to a line parallel to (A_1, A_2). Moreover, $R(A_1A_2A_4A_5)$ is a rectangle with parallel sides, the longest side being parallel to A_1A_2. More generally, this property is true for any $R' \in G$, where G is the group of deplacements generating the tiling.

The additional requirements stated in the lemma come from Lemma 3.1. □

LEMMA 3.3. *Let Ω be a convex polygon with vertices $\{A_0, \dots, A_{p-1}\}$. Then if u is optimal, $u = \max(u_0, \dots, u_{p-1})$, where u_k has the form*

$$u_k(x_1, x_2) = \frac{1}{2\gamma}(X_2^2 - \gamma^2) \quad \text{or} \quad u_k(x_1, x_2) = \cosh\beta\ X_1 + \sinh\beta\ \sqrt{X_1^2 + X_2^2}$$

with $X := Rx + X_0$, where $R \in SO_2$, $X_0 \in \mathbb{R}^2$, and $\gamma, \beta \in \mathbb{R}$ are constant.

Proof. Let A_kA_{k+1} be one of the sides of the polygon. There exists u_k of class C^2 such that $u \equiv u_k$ in a neighborhood V of this side. Also there exists another side A_iA_{i+1} such that for all x, $y(x) \in [A_i, A_{i+1}]$.

There exists $R \in SO_2$, $X_0 \in \mathbb{R}^2$ such that, in the new coordinates system $X = Rx + X_0$, (A_iA_{i+1}) is included in the line $\{X_2 = 0\}$ and either A_kA_{k+1} is parallel to A_iA_{i+1} or the line A_kA_{k+1} contains $X = 0$. Since $u(Y(x)) = 0$, where $Y(x) = Ry(x) + X_0$ and

$$y(x) = x - \tau_x \nabla u(x) = x + 2\frac{u(x)}{1 - |\nabla u(x)|^2}\ \nabla u(x), \tag{20}$$

the fact that $Y_2(x) = 0$ is equivalent to the equation

$$X_2\,(|\nabla u(X)|^2 - 1) = 2u(X)\ \partial_2 u(X). \tag{21}$$

Therefore u_k must be a regular solution of this equation, satisfying $u_k \equiv 0$ on the segment A_kA_{k+1}; that is, either $u(X_1, \gamma) = 0$ for some γ or $u(\delta X_2, X_2) = 0$ for some δ. Since (21) is a Hamilton–Jacobi equation, there is uniqueness of the solution near this line. In the first case, one can check that the solution is $\frac{1}{2\gamma}(X_2^2 - \gamma^2)$, and in the other case, it is

$$\cosh\beta\ X_1 + \sinh\beta\ \sqrt{X_1^2 + xX_2^2},$$

where β is defined by $\sinh\beta = -\delta$. These solutions can be found using the methods described in [5].

Since all these functions are convex, u coincides with them near the sides and satisfies the constraints only if $u \geq U := \max(u_k)$. Moreover, if $u > U$, then $|\nabla u| <$

$|\nabla U|$ in the zone $\{u > U\}$; hence $F(u;\Omega) > F(U;\Omega)$, contradicting the optimality of u. □

In order to conclude the proof of the main theorem, we have to compute explicitly the value of the functional for the polygons given in Lemma 3.2, using the explicit functions given in Lemma 3.3. It turns out that the case of the nonregular hexagons leads to a larger value of the functional. This is shown, with complicated computations, in the appendix.

Let us first present the computation for the rectangle $[-1,1]\times[-k,k]$, where $k \geq 1$ is a given number. We will prove that the value of the functional attains a strict minimum for $k = 1$, that is, for a square.

From Lemma 3.3, since reflection occurs on parallel lines, we have u in the form $\max(u_1,u_2,u_3,u_4)$ with each u_i in the form $u_i(x_1,x_2) = (\pm x_j+\gamma)^2/4\gamma - \gamma$, where $j = 1$ or $j = 2$ and $\gamma = 1$ or $\gamma = k$ in an appropriate coordinate system. If we restrict ourselves to the positive quadrant $x_1 > 0$, $x_2 > 0$, we have

$$u(x_1,x_2) = \max(u_1(x_1),u_2(x_2))$$
$$\text{with} \qquad u_1(x_1) := \frac{(1+x_1)^2}{4} - 1, \quad u_2(x_2) := \frac{(k+x_2)^2}{4k} - k.$$

For any value of $x_1 \in [0,1]$, there exists a unique positive root $x_2 = h(x_1)$ of the equation $u_1(x_1) = u_2(x_2)$, given by

$$h(x_1) = \sqrt{k(4(k-1)+(x_1+1)^2)} - k.$$

Moreover, if $x_2 > h(x_1)$, then $u_2(x_2) > u_1(x_1)$, and hence $u = u_2$; conversely, $u = u_1$ if $x_2 < h(x_1)$. Therefore the value of the functional is given by

$$\begin{aligned} F(u) &= \frac{1}{k}\int_0^1 \left(\int_0^{h(x_1)} f(u_1'(x_1))\,dx_2 + \int_{h(x_1)}^k f(u_2'(x_2))\,dx_2 \right) dx_1 \\ &= 4\int_0^1 \frac{\sqrt{k(4(k-1)+(x_1+1)^2)} - k}{k((x_1+1)^2+4)}\,dx_1 \\ &\quad + \frac{\pi}{2} - 2\int_0^1 \arctan\left[\frac{\sqrt{4(k-1)+(x_1+1)^2}}{2\sqrt{k}}\right] dx_1. \end{aligned}$$

Differentiating with respect to k, using the change of variable $x = (1+x_1)/2$, and writing $k = 1+\beta^2$, we get

$$\frac{dF(u)}{dk} = \frac{\beta^2\sqrt{1+\beta^2}}{2}\int_{\frac{1}{2}}^1 \frac{(x^2-1)^2\,dx}{(1+x^2)(1+2\beta^2+x^2)\sqrt{\beta^2+x^2}} > 0$$

for $\beta \neq 0$; hence $F(u)$ is increasing as claimed.

This proves that the minimum among all rectangles is achieved for a square. The minimal value is given by $k = 1$; that is,

$$F_{\text{square}} := \pi + 12\ln 2 - 4\ln 5 - 4\arctan 2 \simeq 0.5930123.$$

For the regular hexagon, we put the vertices as indicated in Theorem 1. Using symmetries, we have to compute the value of the functional for the triangle OAI,

where $I = (a, 0)$, whose area is $a^2/2\sqrt{3}$. The restriction of u in this triangle is equal to $\phi_a(x_1)$ (defined in Theorem 1); hence the corresponding value of the functional is

$$\begin{aligned} F_{\text{hexagon}} &= \frac{2\sqrt{3}}{a^2} \int_0^a \int_0^{x_1/\sqrt{3}} f(\phi_a'(x_1))\, dx_2\ dx_1 \\ &= \pi + 12 \ln 2 - 4 \ln 5 - 4 \arctan 2 = F_{\text{square}}. \end{aligned}$$

This ends the proof of Theorem 1.

Appendix. Nonregular hexagons. We consider here a hexagon with vertices A, B, C, A', B', C'. We assume that $C'A$ is the longest side of the hexagon; then $C'ACA'$ is a rectangle from Lemma 3.2. We choose a coordinate system where $A = (a, -1)$, $C = (a, 1)$, $B = (a+\alpha, \beta)$ with $\alpha > 0$, and $A' = -A$, $C' = -C$ (see Figure 2). In the following, we consider only the part in $\Omega_+ := \{x \in \Omega;\ x_1 > 0\}$. Indeed, it suffices to minimize the functional on each half part of the hexagon.

We use the auxiliary functions $\phi_a(t) = (t + a)^2/(4a) - a$,

$$\psi_b(x) = x_1 \sqrt{1 + b^2} + b \sqrt{x_1^2 + x_2^2},$$

and the parameters $p := \alpha/(1 + \beta)$, $q := \alpha/(1 - \beta)$. Hence $\alpha = 2pq/(p + q)$, $\beta = (q - p)/(q + p)$. Then the line (AB) crosses $\{x_2 = 1\}$ at $L := (a + 2p, 1)$, and (BC) crosses $\{x_2 = -1\}$ at $K := (a + 2q, -1)$.

Here $u = \max(u_1, u_2, u_3, u_4)$, where

$$\begin{aligned} u_1(x) &= \phi_1(x_2) = \frac{1}{4}(1 + x_2)^2 - 1, \\ u_2(x) &= \phi_1(-x_2) = \frac{1}{4}(1 - x_2)^2 - 1, \\ u_3(x) &= \psi_q(x_1 - a - 2q, 1 + x_2), \\ u_4(x) &= \psi_p(x_1 - a - 2p, 1 - x_2). \end{aligned}$$

We note that $\Omega_i := \text{interior}\{x \in \Omega_+, u(x) = u_i(x)\}$.

THEOREM 3. *Assume that there exists an irregular hexagon H such that $F(u_H, H) < F_{\text{square}}$. Then there exists a symmetrical hexagon H' such that $F(u_{H'}, H') \le F(u_H, H)$.*

Proof. For a symmetrical hexagon, we have $\beta = 0$ or, equivalently, $q = p$. In the following, we assume that $q > p$, and we will prove that the value of F is increasing with respect to q; we note it shortly as $F(q)$:

$$F(q) := \frac{G(q)}{\sum_i |\Omega_i|}, \quad \text{where} \quad G(q) := \sum_{i=1}^4 \int_{\Omega_i} f(\nabla u_i).$$

Let us assume that $F(q)$ is optimal and, in particular, is not greater than F_{square}. Then we have a smaller value if $q = 1/a$. Indeed, we must in general have $qa \ge 1$ since the angle OCB must be not smaller than $\pi/2$ from Lemma 3.1. This implies that the scalar product $OC.CB \ge 0$, that is, $qa \ge 1$, or equivalently the line perpendicular to CB at C must cross the x_1-axis at some point D having a nonnegative first coordinate d. Let $U_1 = \{x \in \Omega_+; |x_1| < d\}$, $U_2 = \Omega_+ \setminus U_1$. If $qa > 1$, then U_1 is nonempty; then

$$F(u; \Omega_+) = tF(u; U_1) + (1 - t)F(u; U_2), \quad \text{where} \quad t := \frac{|U_1|}{|U_1| + |U_2|} \in (0, 1).$$

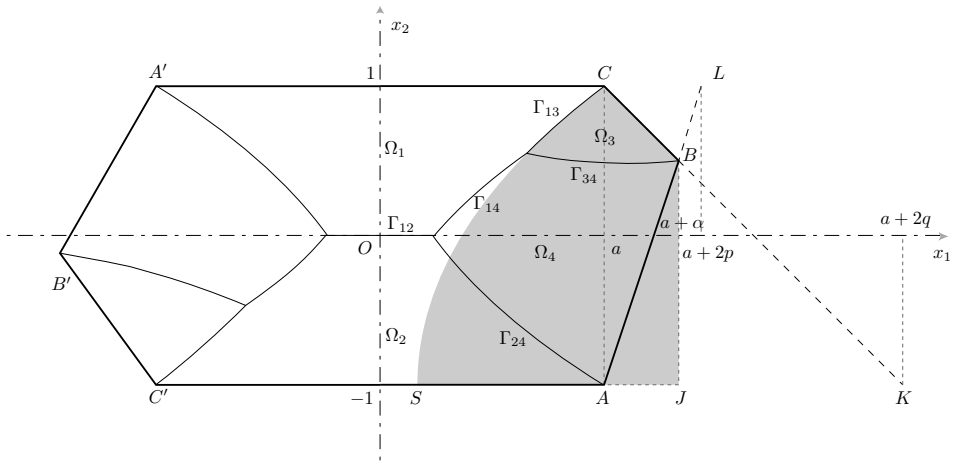


FIG. 2. *Nonsymmetrical hexagon.*

On the other hand, $F(u;U_1) > F_{\text{square}}$ as shown from an explicit computation, using $u = \phi_1(|x_2|)$ in U_1. Therefore, if the optimal value for F is not greater than F_{square}, then $F(u;U_2) \le F_{\text{square}}$; hence F is minimal when we have $U_1 = \emptyset$; that is, $qa = 1$ as claimed.

We are now going to study the variation of $F(q)$ with respect to q, assuming that $qa \equiv 1$. We recall that $F(q) = G(q)/A(q)$, where $A(q) := |\Omega|/2 = 2a + \beta$. We will prove that

$$\frac{\partial A}{\partial q} < 0 \quad \text{and} \quad \frac{\partial G}{\partial q} > 0 \tag{22}$$

under the assumption that $q > p$. (The partial derivatives are understood with p constant.) This implies that F is minimal for $q = p$, that is, for a symmetrical hexagon as claimed.

We note that the angle ABC must be not smaller than $\pi/2$ from Lemma 3.1, that is, $\alpha^2 + \beta^2 \le 1$ or, equivalently, $pq \le 1$. Under the assumption that $p < q$, this implies $p < 1$; since we have $A(q) = 2a + \beta = \frac{2}{q} + \frac{q-p}{q+p}$, we get

$$\frac{\partial A}{\partial q} = -\frac{2}{q^2} + \frac{2p}{(q+p)^2} < -\frac{2}{q^2} + \frac{2}{(q+p)^2} < 0.$$

This proves the first inequality in (22).

Let us turn to the proof of the second one. We note that only u_3 depends on q. Therefore, if we define $g_{ij} = u_i - u_j$, $\Gamma_{ij} = g_{ij}^{-1}(0)$ for $i, j \in \{1, \dots, 4\}$, we have

$$\frac{\partial G}{\partial q} = \int_{\Omega_3} \frac{\partial}{\partial q} f(\nabla u_3) + R_{43} + R_{13},$$

where

$$R_{ij} := \int_{\Gamma_{ij}} \frac{\partial g_{ij}}{\partial q} \, \frac{1}{|\nabla g_{ij}|} \left[f(\nabla u_i) - f(\nabla u_j) \right].$$

We prove that all three terms in $\frac{\partial G}{\partial q}$ are positive. Let us begin with R_{43}.

Let $x \in \Omega_3$ be given, and $y := x - \tau_x \nabla u(x) = x - \tau_x \nabla u_3(x)$. From the definition of u_3, we have $y \in C'A$ and $u(y) = 0$. From (4), we have

$$u_3(x) = -\frac{\tau_x}{2}(1 - |\nabla u_3(x)|^2).$$

A similar relation holds if $x \in \Omega_4$, and $y \in A'C$. For $x \in \Gamma_{43} \subset \overline{\Omega}_3 \cap \overline{\Omega}_4$, there exists τ_3, τ_4 such that

$$-\frac{\tau_3}{2}(1 - |\nabla u_3(x)|^2) = u_3(x) = u_4(x) = -\frac{\tau_4}{2}(1 - |\nabla u_4(x)|^2),$$

and $y_3 := x - \tau_3 \nabla u_3(x) \in C'A$, $y_4 := x - \tau_4 \nabla u_4(x) \in A'C$. Since $\Gamma_{43} \subset \{x_2 > 0\}$, we have $t_3 := |x - y_3| > t_4 := |x - y_4|$; notice also that $t_i = \tau_i |\nabla u_i(x)|$, and, in particular,

$$t_3 \left(\frac{1}{|\nabla u_3(x)|} - |\nabla u_3(x)| \right) = t_4 \left(\frac{1}{|\nabla u_4(x)|} - |\nabla u_4(x)| \right).$$

Since $t_3 > t_4$, this implies that $|\nabla u_3(x)| > |\nabla u_4(x)|$; hence

$$\forall x \in \Gamma_{43}, \qquad f(\nabla u_3(x)) < f(\nabla u_4(x)). \tag{23}$$

Similarly, if $x \in \Gamma_{13}$, y_1 is the orthogonal projection of x on $C'A$; hence $t_1 < t_3$. This implies that

$$\forall x \in \Gamma_{13}, \qquad f(\nabla u_3(x)) < f(\nabla u_1(x)). \tag{24}$$

Moreover, $\frac{\partial}{\partial q} g_{43}(x) = -\frac{\partial}{\partial q} u_3(x) = \frac{\partial}{\partial q} g_{13}(x)$ since $u_4(x)$ and $u_1(x)$ do not depend on q. In what follows, we will use the following coordinates:

$$X_1 := a + 2q - x_1, \qquad X_2 = 1 + x_2, \qquad R := \sqrt{X_1^2 + X_2^2}.$$

We recall that $x \in \overline{\Omega}_3$ in the expression of $\partial G/\partial q$. We need to describe the domain of variation of (X_1, R) for $x \in \overline{\Omega}_3$.

We have

$$\overline{\Omega}_3 \subset \{x; \quad u_3(x) \geq u_1(x) \text{ and } u_3(x) \leq 0 \text{ and } x_1 \leq a + \alpha \text{ and } x_2 \geq -1\};$$

that is, $\overline{\Omega}_3$ is included in the gray zone $SJBC$ indicated on Figure 2. This yields first $R \geq X_1$. Then $x_1 \leq a + \alpha$ implies $X_1 \geq 2q^2/(q + p) \geq q$ since we assume $q > p$. The condition $u_3 \leq 0$ implies $qR \leq X_1\sqrt{1 + q^2}$. And $u_3 \geq u_1$ leads to $(X_1 - 2\sqrt{1 + q^2})^2 \geq (R - 2q)^2$; taking into account that $X_1 \leq R$, this yields $R + X_1 \leq 2(q + \sqrt{1 + q^2})$, and, in particular, $X_1 \leq q + \sqrt{1 + q^2}$. Therefore, in (X_1, R) coordinates, $\overline{\Omega}_3$ is included in the polygon $\widetilde{B}\widetilde{C}\widetilde{S}\widetilde{J}$ pictured in Figure 3.

We have

$$\begin{aligned} -\frac{\partial}{\partial q} u_3(x) &= -\frac{\partial}{\partial q} \psi_q(a + 2q - x_1, 1 + x_2) \\ &= -\frac{\partial \psi_q}{\partial q}(a + 2q - x_1, 1 + x_2) - 2\frac{\partial \psi_q}{\partial x_1}(a + 2q - x_1, 1 + x_2) \\ &= \frac{-R^2 + R\,(X_1 q + 2(1 + q^2)) - 2qX_1\sqrt{1 + q^2}}{R\sqrt{1 + q^2}}. \end{aligned}$$

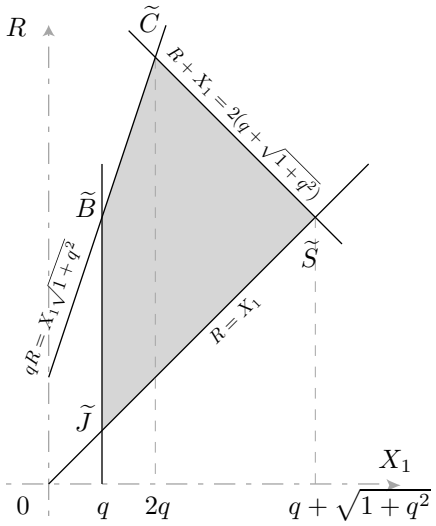


FIG. 3. *Zone including* $\overline{\Omega}_3$ *in* (X_1, R) *coordinates..*

Hence the sign of $\frac{\partial}{\partial q}u_3(x)$ is the same as the sign of

$$n_3(X_1, R) := -R^2 + R\,(X_1 q + 2(1+q^2)) - 2qX_1\sqrt{1+q^2}.$$

This is a concave expression in R, and so it is minimal for the extremal values of R, when X_1 is given. That is, we have to check that $n_3 \geq 0$ on $[\widetilde{J}\widetilde{S}]$, $[\widetilde{C}\widetilde{S}]$, $[\widetilde{B}\widetilde{C}]$.

On $[\widetilde{J}\widetilde{S}]$, we have $R = X_1$; using $m := q + \sqrt{1+q^2}$, we get

$$\begin{aligned} n_3(X_1, X_1) &= X_1(X_1(q-1) + 2(1+q^2) - 2q\sqrt{1+q^2}) \\ &= X_1 m^{-2}(-X_1 m + m^2 + 1) \geq 0, \end{aligned}$$

since $X_1 \leq m$.

On $[\widetilde{C}\widetilde{S}]$, we have $R = 2m - X_1$, and

$$n_3(X_1, 2m - X_1) = -m^4 + 2X_1 m^3 - X_1^2 m^2 + 1.$$

This is again a concave expression on X_1, so it is minimal for $X = m$ or $X = 2q = m - 1/m$; since $n_3(m, m) = 1$ and $n_3(m, m + 1/m) = 0$, we conclude that $n_3 \geq 0$ on $[\widetilde{C}\widetilde{S}]$.

On $[\widetilde{B}\widetilde{C}]$, $q \leq X_1 \leq 2q$, $R = X_1\sqrt{1+q^2}/q$; hence

$$n_3(X_1, X_1\sqrt{1+q^2}/q) = -2mX_1(1+m^2)\frac{X_1 m - m^2 + 1}{(m^2-1)^2} \geq 0.$$

This proves that $n_3 \geq 0$; hence $\frac{\partial}{\partial q}g_{43}(x) = \frac{\partial}{\partial q}g_{13}(x) \geq 0$. This proves that $R_{43} > 0$ and $R_{13} > 0$, taking (23) and (24) into account.

To conclude the proof that $\partial G/\partial q > 0$, we prove that $\frac{\partial}{\partial q}f(\nabla u_3) > 0$ or, equivalently, that $\frac{\partial}{\partial q}\left|\nabla u_3\right|^2 < 0$. We have

$$\begin{aligned} \left|\nabla u_3\right|^2 = &\left[\frac{q(a+2q-x_1)}{\sqrt{(a+2q-x_1)^2 + (1+x_2)^2}} - \sqrt{1+q^2}\right]^2 \\ &+ \frac{q^2(1+x_2)^2}{(a+2q-x_1)^2 + (1+x_2)^2}. \end{aligned}$$

Hence, using the notations X_1, X_2, R,

$$\frac{\partial}{\partial q}\left|\nabla u_3\right|^2 = \frac{2N(X_1,R)}{R^3\sqrt{1+q^2}},$$

where

$$N(X_1,R) := 2qX_1^2(1+q^2) - (2q^2+1)R^2X_1 - 2q(1+q^2)R^2 + 2q\sqrt{1+q^2}R^3.$$

We have to prove that $N \leq 0$; since this is a convex function of X_1, it is enough to prove that for the extremal values of X_1.

Let us first examine the value on $[\widetilde{S}\widetilde{J}]$, where $X_1 = R$. We have

$$N(R,R) = R^3(2q\sqrt{1+q^2} - 2q^2 - 1) < 0$$

since $2q\sqrt{1+q^2} < 2q^2+1$ for any $q > 0$.

On $[\widetilde{B}\widetilde{J}]$, we have $X_1 = q$, and

$$N(q,R) := 2q\sqrt{1+q^2}R^3 - (4q^2+3)qR^2 + 2q^3(1+q^2).$$

Since $\frac{d}{dR}N(q,R)$ has the form $R(c_1R - c_2)$, with $c_1, c_2 > 0$, $N(q,R)$ is decreasing, then increasing on $\mathbb{R}_+$. Hence, to prove $N(q,R) < 0$, it is enough to check on the boundary values for R, that is, $R = q$ (for $\widetilde{J}$) and $R = \sqrt{1+q^2}$ (for $\widetilde{B}$). We already know that $N(q,q) < 0$ and

$$N(q,\sqrt{1+q^2}) = -q(1+q^2) < 0.$$

On $[\widetilde{B}\widetilde{C}]$, we have $R = X_1\sqrt{1+q^2}/q$ and then

$$N(X_1, X_1\sqrt{1+q^2}/q) := \frac{X_1^2}{q^2}(1+q^2)[X_1 - 2q] \leq 0,$$

since $X_1 \leq 2q$ on $[\widetilde{B}\widetilde{C}]$.

On $[\widetilde{C}\widetilde{S}]$, we have $R = 2m - X_1$, where $m := q + \sqrt{1+q^2}$ and $X_1 \in [2q, m]$. Using $t := m(X - 2q)$, we have $t \in [0,1]$; thus,

$$N(X_1, 2m - X_1) = N\left(2q + \frac{t}{m}, 2m - 2q - \frac{t}{m}\right) = \frac{t}{m^3}\widetilde{N}(t),$$

where

$$\widetilde{N}(t) = -m^2t^2 + (2m^4 + 3m^2 - 1)t - 3m^4 - 2m^2 + 1.$$

We have $\widetilde{N}'(t) = -2tm^2 + 2m^4 + 3m^2 - 1 > 0$ since $m \geq 1$ and $t \in [0,1]$. Hence

$$\widetilde{N}(t) \leq \widetilde{N}(1) = -m^4 < 0.$$

This concludes the proof of (22) and the proof of Theorem 3. □

We now consider the case of a symmetrical hexagon, that is, $\beta = 0$, $p = q = \alpha$. For the reason given in the beginning of the proof of Theorem 3, it is sufficient to consider the case $a = 1/\alpha$, with $a \geq 1$.

THEOREM 4. *Let $F(a)$ be the value of $F(u_{H_a}; H_a)$, where H_a is the symmetrical hexagon ($\beta = 0$) with parameters $a \geq 1$, $\alpha = 1/a$. Then $F(a)$ is increasing with respect to $a \geq 1$, and its minimal value $F(1)$ is greater than F_{square}.*

Proof. We prove the theorem by an explicit computation of $F(a)$ and then by a differentiation with respect to a.

Since $\beta = 0$, the point B in Figure 2 lies on the x_1-axis, and H_a and u_{H_a} are symmetrical with respect to x_1 and x_2. Therefore, if we define $\Omega_a := H_a \cap \{x_1 > 0, x_2 > 0\}$ and u_a as the restriction of u_{H_a} to Ω_a, we have $F(a) = F(u_a, \Omega_a)$. In Ω_a, we have $u_a = \max(u_1, u_3)$, where $u_1(x) = \phi_1(x_2)$ and $u_3(x) = \psi_{1/a}(x_1 - a - 2/a, 1 + x_2)$.

We have $\overline{\Omega}_a = \overline{\Omega}_1 \cup \overline{\Omega}_3$ with $u_a = u_i$ in Ω_i, $i = 1, 3$. We still note by Γ_{13} the common boundary of Ω_1 and Ω_3. It is defined by the equation

$$\begin{aligned}\frac{1}{4}(1 + x_2)^2 - 1 = u_1(x) &= u_3(x) \\ &= \sqrt{1 + a^{-2}} \left(x_1 - a - \frac{2}{a} \right) + \frac{\sqrt{(x_1 - a - 2/a)^2 + (1 + x_2)^2}}{a},\end{aligned}$$

which can be solved in the form

$$x_1 = w(x_2) = \frac{(s - 1)(2s^2 + 2 + (x_2 + 1)^2 s)}{4s(s + 1)},$$

where $s := \exp(\operatorname{asinh} a) = a + \sqrt{1 + a^2}$.

We have

$$F(a) = \frac{G_1(a) + G_3(a)}{|\Omega_a|}, \quad \text{where} \quad G_i(a) := \int_{\Omega_i} f(\nabla u_i). \tag{25}$$

Now

$$G_3(a) = \int_0^1 \int_{w(x_2)}^{a + \frac{1 - x_2}{a}} f_3(x_1, x_2)\, dx_1\, dx_2$$

$$\text{with} \quad f_3(x_1, x_2) := f(\nabla u_3(x)) = \frac{1}{2} \frac{a^2 \sqrt{X^2 + Y^2}}{\sqrt{1 + a^2} X + (1 + a^2)\sqrt{X^2 + Y^2}},$$

where $X := x_1 - a - 2/a$, $Y := x_2 + 1$.

Using the change of variable $X = Y \sinh t$, we have that

$$J_3 := \int_{w(x_2)}^{a + \frac{1 - x_2}{a}} f_3(x_1, x_2)\, dx_1 = \frac{a^2 Y}{2} \int_W^{-\operatorname{asinh} \frac{1}{a}} \frac{\cosh^2 t\, dt}{(a^2 + 1)\cosh t + \sqrt{1 + a^2} \sinh t},$$

where $\sinh W = (aw(Y - 1) - 2 - a^2)/(aY)$, that is,

$$W = \log \left(\frac{Y(s - 1)}{2(s + 1)} \right),$$

taking into account the value of w and using $s = a + \sqrt{1 + a^2}$. We explicitly get

$$\begin{aligned}8(s^4 - 1) J_3 = (1 + s^4 + 6s^2)(4 - Y^2) + 4s(s^2 + 1)(Y - 2)^2 \\ + 32 s^2 Y \arctan \left(\frac{Y - 2}{2 + Y} \right).\end{aligned}$$

Integrating for $Y \in [1,2]$, we get

$$G_3(a) = \frac{5s^4 + 4s^3 + (144\arctan 3 + 384\arctan 2 - 66 - 168\pi)s^2 + 4s + 5}{24(s^4-1)}.$$

Since u_1 does not depend on x_1, we have

$$\begin{aligned} G_1(a) &= \int_0^1 w(x_2) f(\nabla u_1(x_2))\, dx_2 \\ &= \frac{(s-1)((4\arctan 2 - \pi)(s-1)^2 + 4s)}{4(s+1)s}. \end{aligned}$$

Finally, $|\Omega_a| = a + 1/(2a) = (1+s^4)/(2s(s^2-1))$, so

$$F(a) = \frac{2s(s^2-1)}{1+s^4}\left[G_1(a) + G_3(a)\right].$$

Hence,

$$F'(a) = \frac{ds}{da}\ \frac{(s^2-1)^2 N(s)}{12(1+s^4)^2(s^2+1)^2}$$

has the sign of

$$\begin{aligned} N(s) := 96\left(1 - 9s^5 - 6s^4 - 9s^3 - 6s^6 - 6s^2 - 3s - 3s^7 + s^8\right)\arctan 2 \\ - 144s^2(3 + 4s^2 + 3s^4)\arctan 3 \\ - 24(1 - 18s^4 - 9s^5 - 9s^3 - 15s^2 - 3s^7 + s^8 - 3s - 15s^6)\pi \\ - 29 + 88s + 54s^2 + 264s^3 + 88s^7 + 14s^4 + 264s^5 - 29s^8 + 54s^6. \end{aligned}$$

Since $a \geq 1$, $s \geq 1 + \sqrt{2}$. Computing the coefficients explicitly, we get for $t > 0$,

$$\begin{aligned} N(1+\sqrt{2}+t) \simeq 1.88805320\, t^8 + 31.8011496\, t^7 + 236.968636\, t^6 \\ + 1013.948295\, t^5 + 2708.11097\, t^4 + 4589.26377\, t^3 \\ + 4768.4986\, t^2 + 2727.7799\, t + 634.0078 > 0. \end{aligned}$$

Hence $F(a)$ is increasing as claimed. The minimal value of $F(a)$ is

$$\begin{aligned} F(1) &= ((528\sqrt{2} + 768)\arctan 2 + (180\sqrt{2} + 252)\arctan 3 \\ &\quad - (222\sqrt{2} + 318)\pi + 53 + 41\sqrt{2}) \Big/ (36(10 + 7\sqrt{2})) \\ &\simeq 0.60771279 > F_{\text{square}}. \end{aligned}$$

This concludes the proof of the theorem. □

REFERENCES

[1] M. Berger, *Géométrie*, Nathan, Paris, 1990.

[2] G. Buttazzo, V. Ferone, and B. Kawohl, *Minimum problems over sets of concave functions and related questions*, Math. Nach., 173 (1993), pp. 71–89.

[3] M. Comte and T. Lachand-Robert, *Existence of minimizers for Newton's problem of the body of minimal resistance under a single impact assumption*, J. Anal. Math., 83 (2001), pp. 313–335.

[4] M. Comte and T. Lachand-Robert, *Newton's problem of the body of minimal resistance under a single-impact assumption*, Calc. Var. Partial Differential Equations, 12 (2001), pp. 173–211.

[5] G. M. Murphy, *Ordinary Differential Equations and Their Solutions*, D. Van Nostrand, Princeton, NJ, 1960.

[6] I. Newton, *Philosophiae Naturalis Principia Mathematica*, 1686.

SIAM J. MATH. ANAL.
Vol. 34, No. 1, pp. 121–141

ON THE CAUCHY PROBLEM FOR STOCHASTIC STOKES EQUATIONS*

R. MIKULEVICIUS†

Abstract. We extend Krylov's L_p-solvability theory of the second order quasi-linear parabolic stochastic differential equations to the stochastic Stokes equation. Some additional integrability and regularity properties are also presented.

Key words. stochastic partial differential equations, Stokes equation, Cauchy problem

AMS subject classifications. 60H15, 35R60

PII. S0036141001390312

1. Introduction.

In this paper we study the stochastic Stokes equation. Specifically, we are considering in $\mathbf{R}^d$ the system of equations for $\mathbf{u}=(u^l)_{1\le l\le d}$ and scalar functions p, $\tilde{p}$:

$$\begin{aligned} \partial_t u^l &= \partial_i(a^{ij}(t,x)\partial_j u^l)+D^l(\mathbf{u},t,x)+\partial_l p \\ &\quad +[\sigma^k(t,x)\partial_k u^l+Q^l(\mathbf{u},t,x)+\partial_l\tilde{p}]\cdot\dot{W}, \\ u^l(0,x) &= u_0^l(x), \quad l=1,\dots,d, \quad x\in\mathbf{R}^d, \\ \operatorname{div}\mathbf{u} &= 0, \end{aligned} \tag{1.1}$$

where W is a cylindrical Wiener process in a Hilbert space. Here and everywhere below, the summation with respect to the repeated indices is assumed.

Our interest was motivated by stochastic fluid mechanics (see, e.g., [3], [4]). While the assumptions imposed below exclude the "typical" nonlinearity $u^k\partial_k u$, they allow us to construct suitable approximations to the solution of stochastic Navier–Stokes equations.

In [1], [2], Krylov developed a comprehensive theory of second order quasi-linear parabolic stochastic differential equations in Bessel classes $H_p^s(\mathbf{R}^d)$. In [7], Krylov's results were extended to parabolic systems of quasi-linear stochastic PDEs on $\mathbf{R}^d$ for $\mathbf{u}=(u^l)_{1\le l\le d}$:

$$\begin{aligned} \partial_t u^l &= \partial_i(a^{ij}(t,x)\partial_j u^l)+D^l(\mathbf{u},t,x) \\ &\quad +[\sigma^k(t,x)\partial_k u^l+Q^l(\mathbf{u},t,x)]\cdot\dot{W}, \\ u^l(0,x) &= u_0^l(x), \quad l=1,\dots,d, \quad x\in\mathbf{R}^d. \end{aligned}$$

We prove the existence and uniqueness of solutions to (1.1) in the spaces of Bessel potentials.

The structure of the paper is as follows. In section 2 we introduce the notation and state the main result about the existence and uniqueness of solutions to (1.1). In section 3 we present some auxiliary results needed for the investigation of the Stokes equation in $\mathbb{R}^d$. They regard pointwise multipliers in $\mathbf{R}^d$ and solenoidal and potential

*Received by the editors June 4, 2001; accepted for publication (in revised form) January 23, 2002; published electronically August 15, 2002. The author was supported by NSF grant DMS-98-02423.
http://www.siam.org/journals/sima/34-1/39031.html

†Institute of Mathematics and Informatics, Akademijos 4, Vilnius, Lithuania (mikulvcs@math.usc.edu).

projections of vector fields. In section 4, following Krylov's ideas and [7], we prove the main results about the existence and uniqueness of solutions to (1.1). Also, some additional integrability and regularity properties of solutions are investigated.

2. Notation and the main results. Let Y be a separable Hilbert space with a norm $|\cdot|_Y$. The scalar product of $x, y \in Y$ will be denoted by $x \cdot y$.

Let $\mathbf{R}^d$ be a d-dimensional Euclidean space with elements $x = (x_1, \dots, x_d)$; if $x, y \in \mathbf{R}^d$, we write

$$(x, y) = \sum_{i=1}^{d} x_i y_i, \qquad |x| = \sqrt{(x, x)}.$$

If u is a function on $\mathbf{R}^d$, the following notational conventions will be used for its partial derivatives: $\partial_i u = \partial u/\partial x_i$, $\partial_{ij}^2 = \partial^2 u/\partial x_i \partial x_j$, $\partial_t u = \partial u/\partial t$, $\nabla u = \partial u = (\partial_1 u, \dots, \partial_d u)$, and $\partial^2 u = (\partial_{ij}^2 u)$ denotes the Hessian matrix of second derivatives. Let $\alpha = (\alpha_1, \dots, \alpha_d)$ be a multi-index; then $\partial_x^\alpha = \Pi_{i=1}^d \partial_{x_i}^{\alpha_i}$.

Let $C_0^\infty = C_0^\infty(\mathbf{R}^d)$ be the set of all infinitely differentiable functions on $\mathbf{R}^d$ with compact support.

For $s \in (-\infty, \infty)$, write $\Lambda^s = \Lambda_x^s = (1 - \sum_{i=1}^d \partial^2/\partial x_i^2)^{s/2}$.

For $p \in [1, \infty]$ and $s \in (-\infty, \infty)$, we define the space $H_p^s = H_p^s(\mathbf{R}^d)$ as the space of generalized functions u with the finite norm

$$|u|_{s,p} = |\Lambda^s u|_p,$$

where $|\cdot|_p$ is the L_p norm. Obviously, $H_p^0 = L_p$. Note that if $s \geq 0$ is an integer, the space H_p^s coincides with the Sobolev space $W_p^s = W_p^s(\mathbf{R}^d)$.

If $p \in [1, \infty]$ and $s \in (-\infty, \infty)$, $H_p^s(Y) = H_p^s(\mathbf{R}^d, Y)$ denotes the space of Y-valued functions on $\mathbf{R}^d$ so that the norm $||g||_{s,p} = |\,|\Lambda^s g|_Y\,|_p < \infty$. We also write $L_p(Y) = L_p(\mathbf{R}^d, Y) = H_p^0(Y) = H_p^0(\mathbf{R}^d, Y)$. Let $C_0^\infty(Y)$ be the space of Y-valued infinitely differentiable functions on $\mathbf{R}^d$ with compact support.

Obviously, the spaces $C_0^\infty, C_0^\infty(Y)$, $H_p^s(\mathbf{R}^d)$, and $H_p^s(\mathbf{R}^d, Y)$ can be extended to vector functions (denoted with boldfaced letters). For example, the space of all vector functions $\mathbf{u} = (u^1, \dots, u^d)$ such that $\Lambda^s u^l \in L_p$, $l = 1, \dots, d$, with the finite norm

$$|\mathbf{u}|_{s,p} = \left(\sum_l |u^l|_{s,p}^p \right)^{1/p},$$

we denote by $\mathbb{H}_p^s = \mathbb{H}_p^s(\mathbf{R}^d)$. Similarly, we denote by $\mathbb{H}_p^s(Y) = \mathbb{H}_p^s(\mathbf{R}^d, Y)$ the space of all vector functions $g = (g^l)_{1 \leq l \leq d}$, with Y-valued components g^l, $1 \leq l \leq d$, so that $||g||_{s,p} = (\sum_l |g^l|_{s,p}^p)^{1/p} < \infty$. The set of all infinitely differentiable vector functions $u = (u^1, \dots, u^d)$ on $\mathbf{R}^d$ with compact support will be denoted by $\mathbb{C}_0^\infty$. We denote by $\mathbb{C}_0^\infty(Y)$ the set of all infinitely differentiable vector functions $u = (u^1, \dots, u^d)$ on $\mathbf{R}^d$ with compact support. (All u^l are Y-valued.)

When $s = 0$, $\mathbb{H}_p^s(Y) = \mathbb{L}_p(Y) = \mathbb{L}_p(\mathbf{R}^d, Y)$. Also, in this case, the norm $||g||_{0,p}$ is denoted more briefly by $||g||_p$. To forcefully distinguish L_p norms in spaces of Y-valued functions, we write $||\cdot||_p$, while in all other cases a norm is denoted by $|\cdot|$.

The duality $\langle \cdot, \cdot \rangle_s$ between $\mathbb{H}_p^s(\mathbf{R}^d, Y)$ and $\mathbb{H}_q^{-s}(\mathbf{R}^d, Y)$, $p \geq 2$, $s \in (-\infty, \infty)$, and

$q = p/(p-1)$, is defined by

$$\langle \boldsymbol{\phi}, \boldsymbol{\psi} \rangle_{\mathbb{H}^s_p(Y),\mathbb{H}^{-s}_q(Y)} = \sum_{i=1}^{d} \int_{\mathbf{R}^d} \left[\Lambda^s \phi^i\right](x) \cdot \Lambda^{-s}\psi^i(x)\, dx$$

$$= \sum_{i=1}^{d} \int_{\mathbf{R}^d} \left[\tilde{\Lambda}_s \phi^i\right](x) \cdot \tilde{\Lambda}_{-s}\psi^i(x)\, dx, \quad \boldsymbol{\phi} \in \mathbb{H}^s_p(Y), \quad \boldsymbol{\psi} \in \mathbb{H}^{-s}_q(Y).$$

If $Y = \mathbf{R}$, i.e., $\mathbb{H}^s_p(Y) = \mathbb{H}^s_p$, we denote $\langle \boldsymbol{\phi}, \boldsymbol{\psi} \rangle_{\mathbb{H}^s_p(Y),\mathbb{H}^{-s}_q(Y)} = \langle \boldsymbol{\phi}, \boldsymbol{\psi} \rangle_s = \langle \boldsymbol{\phi}, \boldsymbol{\psi} \rangle_{s,p}$. If $\mathbf{f} \in \mathbb{H}^s_q(\mathbf{R}^d, Y)$ and $\boldsymbol{\phi} \in \mathbb{H}^{-s}_p(\mathbf{R}^d)$, $p \geq 2$, $s \in (-\infty, \infty)$, and $q = p/(p-1)$, we write

$$\langle \mathbf{f}, \boldsymbol{\phi} \rangle_{s,Y} = \langle \mathbf{f}, \boldsymbol{\phi} \rangle_{s,p,Y} = \sum_{l=1}^{d} \int_{\mathbf{R}^d} \left[\Lambda^s f^l(x)\right] \Lambda^{-s}\phi^l(x)\, dx$$

$$= \sum_{l=1}^{d} \int_{\mathbf{R}^d} \left[\tilde{\Lambda}_s f^l(x)\right] \tilde{\Lambda}_{-s}\phi^l(x)\, dx.$$

Obviously, the function $\boldsymbol{\phi} \longrightarrow \langle \mathbf{f}, \boldsymbol{\phi} \rangle_{s,Y}$ is a linear mapping from H^{-s}_p into Y, and $|\langle \mathbf{f}, \boldsymbol{\phi} \rangle_s|_Y \leq ||\mathbf{f}||_{s,q} |\boldsymbol{\phi}|_{-s,p}$. Similar notation, $\langle \phi, \psi \rangle_s$ and $\langle f, \phi \rangle_{s,Y}$, will be used for scalar functions.

We define the subspace of the divergence free vector fields $\mathcal{S}(\mathbb{H}^s_p(Y)) = \{\mathbf{v} \in \mathbb{H}^s_p(Y) : \operatorname{div} \mathbf{v} = 0\} \subseteq \mathbb{H}^s_p(Y)$ and the subspace of gradient vector fields

$$\mathcal{G}(\mathbb{H}^s_p(Y)) = \{\mathbf{v} \in \mathbb{H}^s_p(Y) : \langle \mathbf{v}, \mathbf{g} \rangle_{\mathbb{H}^s_p(Y),\mathbb{H}^{-s}_q(Y)} = 0 \ \forall\ \mathbf{g} \in \mathcal{S}(\mathbb{H}^{-s}_q(Y))\},$$

where $p \geq 2$, $q = p/(p-1)$, $s \in (-\infty, \infty)$.

Also, we will need some spaces of Y-valued continuous functions. For $m = 1, 2, 3, \ldots$, we define

$$C^m(Y) = \{u : \partial^\alpha u \text{ is uniformly continuous on } \mathbf{R}^d \ \forall\ |\alpha| \leq m\},$$

with the norm $||u||_{C^m} = \sum_{|\alpha| \leq m} ||\partial^\alpha u||_\infty$. For a noninteger $s > 0$, we define

$$\mathcal{C}^s(Y) = \left\{ u \in C^{[s]} : ||u||_{\mathcal{C}^s} = ||u||_{C^{[s]}} + \sum_{|\alpha|=[s]} \sup_{x \neq y} \frac{|\partial^\alpha u(x) - \partial^\alpha u(y)|_Y}{|x-y|^{\{s\}}} < \infty \right\},$$

where $s = [s] + \{s\}$, s is an integer, and $0 \leq \{s\} < 1$. For an integer $s > 0$, we denote

$$\mathcal{C}^s(Y) = \left\{ u \in C^{s-1} : ||u||_{\mathcal{C}^s} = ||u||_{C^{s-1}} + \sum_{|\alpha|=[s]^-} \sup_{x \neq y} \frac{|\partial^\alpha u(x) - \partial^\alpha u(y)|_Y}{|x-y|} < \infty \right\},$$

where $s = [s]^- + 1$. If $Y = \mathbf{R}^d$, we write simply $C^m, \mathcal{C}^s$.

Remark 2.1 (*see, for example, Lemma* 6 *in* [7]). Let $s > 0$. Then

(a) $H^s_\infty(Y) \subseteq \mathcal{C}^s(Y)$, if s is not an integer;

(b) $\mathcal{C}^{s+\varepsilon}(Y) \subseteq H^s_\infty(Y)$ for each $\varepsilon > 0$.

Let

$$B^s(Y) = \begin{cases} H^s_\infty(Y) & \text{if } s > 0 \text{ is not an integer,} \\ \mathcal{C}^s(Y) & \text{if } s > 0 \text{ is an integer,} \\ L_\infty(Y) & \text{if } s = 0, \end{cases}$$

and denote the corresponding norms by $|\cdot|_{B^s}$. If $Y=\mathbf{R}^d$, we write simply B^s.

It is shown (see Lemma 7 in [7]) that for $b\in B^{|s|}(Y)$, $s\in(-\infty,\infty)$, $p\in(1,\infty)$, there is a constant N so that

$$||b\mathbf{v}||_{s,p}\le N||b||_{B^{|s|}}|\mathbf{v}|_{s,p} \tag{2.1}$$

for all $\mathbf{v}\in\mathbb{H}^s_p$, where $b\mathbf{v}=(bv^1,\dots,bv^d)$.

Let $(\Omega,\mathcal{F},\mathbf{P})$ be a probability space with a filtration $\mathbb{F}$ of right continuous σ-algebras $(\mathcal{F}_t)_{t\ge0}$. All the σ-algebras are assumed to be $\mathbf{P}$-completed. Let $W(t)$ be an $\mathbb{F}$-adapted cylindrical Brownian motion in Y; i.e., we have a family of continuous martingales $W_t(v)$, $v\in Y$, with the quadratic variation

$$\langle W(v),W(v')\rangle_t=t\ v\cdot v'\qquad\forall\, v,\,v'\in Y.$$

Let $p\ge2$, $s\in(-\infty,\infty)$. Denote by $\mathcal{I}_{s,p}$ the set of all measurable $\mathbb{F}$-adapted $\mathbb{H}^s_p(Y)$-valued functions such that for every t,

$$\int_0^t||\mathbf{g}(r)||^p_{s,p}\,dr<\infty,\quad\mathbf{P}\text{ a.s.}$$

It is shown (see Theorem 2.1 in [7]) that for each $\mathbf{g}\in\mathcal{I}_{s,p}$ there is a unique $\mathbb{H}^s_p(Y)$-valued continuous martingale $\mathbf{M}(t)=\int_0^t\mathbf{g}(r)\cdot dW(r)$ such that for all $\boldsymbol{\phi}\in\mathbb{H}^{-s}_q$,

$$\left\langle\int_0^t\mathbf{g}(r)\cdot dW(r),\boldsymbol{\phi}\right\rangle_s=\int_0^t\langle\mathbf{g}(r),\boldsymbol{\phi}\rangle_{s,Y}\cdot dW(r)\quad\forall t>0,\mathbf{P}\text{ a.s.}$$

Moreover, for each $T>0$ there exists a constant C not depending on g so that for any stopping time $\tau\le T$,

$$\mathbf{E}\sup_{r\le\tau}|\mathbf{M}(r)|^p_{s,p}\le C\mathbf{E}\int_0^\tau||\mathbf{g}(r)||^p_{s,p}\,dr. \tag{2.2}$$

Consider the Stokes system (1.1) in vector form:

$$\begin{aligned}\partial_t\mathbf{u}(t,x)&=\partial_i(a^{ij}(t,x)\partial_j\mathbf{u})+\mathbf{D}(\mathbf{u},t,x)-\nabla p(t,x)\\&\qquad[\sigma^k(t,x)\partial_k\mathbf{u}(t,x)+\mathbf{Q}(\mathbf{u},t,x)-\nabla\tilde p(t,x)]\cdot\dot W,\\\mathbf{u}(0,x)&=\mathbf{u}_0(x),\qquad\operatorname{div}\mathbf{u}=0.\end{aligned} \tag{2.3}$$

Let $s\in(-\infty,\infty)$, $p\ge2$. For $\mathbf{v}\in\mathbb{H}^{s+1}_p$, let $\mathbf{Q}(\mathbf{v},t)=\mathbf{Q}(\mathbf{v},t,x)$ be a predictable $\mathbb{H}^s_p(Y)$-valued function and $\mathbf{D}(\mathbf{v},t)=\mathbf{D}(\mathbf{v},t,x)$ a predictable $\mathbb{H}^{s-1}_p$-valued function. Let $a=a(t)=(a^{ij}(t,x))_{1\le i,j\le d}$ be a symmetric $\mathbb{F}$-adapted matrix. Let $\sigma=\sigma(t)=(\sigma^k(t,x))_{1\le k\le d}$ be an $\mathbb{F}$-adapted vector function with Y-valued components σ^k, and let $\mathbf{u}_0=(u^l_0)_{1\le l\le d}$ be an $\mathcal{F}_0$-measurable $\mathbb{H}^{s+1-2/p}_p$-valued function so that $\mathbf{E}||\mathbf{u}_0||^p_{s+1-2/p,p}<\infty$, $\operatorname{div}\mathbf{u}_0=0$.

The following assumptions will be made:

A. For all $t\ge0$, $x,\lambda\in R^d$,

$$K|\lambda|^2\ge\left[a^{ij}(t,x)-\frac12\sigma^i(t,x)\cdot\sigma^j(t,x)\right]\lambda^i\lambda^j\ge\delta|\lambda|^2,$$

where K,δ are fixed strictly positive constants.

A1(s,p). For all $t,x,y,\mathbf{P}$ a.s.,

$$|a^{ij}(t,x)-a^{ij}(t,y)|+|\sigma^i(t,x)-\sigma^i(t,y)|_Y\le K|x-y|$$

and

$$\begin{cases} |a^{ij}(t)|_{B^s}\le K & \text{if } s\ge 1,\\ |a(t,x)|\le K & \text{if } -1<s<1,\\ |a^{ij}(t)|_{B^{-s+\varepsilon}}\le K & \text{if } s\le -1, \end{cases}$$

where $\varepsilon\in(0,1)$.

For all i,t,x,

$$\begin{cases} ||\sigma^i(t)||_{B^s}\le K & \text{if } s\ge 1,\\ |\sigma^i(t,x)|_Y\le K & \text{if } s\in(-1,1),\\ ||\sigma^i(t)||_{B^{-s+\varepsilon}}\le K & \text{if } s\le -1, \end{cases}$$

where $\varepsilon\in(0,1)$.

A2(s,p). For $\mathbf{v}\in\mathbb{H}^{s+1}_p$, $\mathbf{Q}(\mathbf{v},t)=\mathbf{Q}(\mathbf{v},t,x)$ is a predictable $\mathbb{H}^s_p(Y)$-valued function and $\mathbf{D}(\mathbf{v},t)=\mathbf{D}(\mathbf{v},t,x)$ is a predictable $\mathbb{H}^{s-1}_p$-valued function, and $\mathbf{P}$ a.s. for each t

$$\int_0^t(|\mathbf{D}(\mathbf{0},r)|^p_{s-1,p}+||\mathbf{Q}(\mathbf{0},r)||^p_{s,p})\,dr<\infty\quad\forall t>0,\mathbf{P}\text{ a.s.},$$

where $\mathbf{0}=(0,\dots,0)$.

A3(s,p). For every $\varepsilon>0$, there exists a constant K_ε such that for any $\mathbf{u},\mathbf{v}\in\mathbb{H}^{s+1}_p$,

$$|\mathbf{D}(\mathbf{u},t,x)-\mathbf{D}(\mathbf{v},t,x)|_{s-1,p}+||\mathbf{Q}(\mathbf{u},t,x)-\mathbf{Q}(\mathbf{v},t,x)||_{s,p}$$

$$\le\varepsilon|\mathbf{u}-\mathbf{v}|_{s+1,p}+K_\varepsilon|\mathbf{u}-\mathbf{v}|_{s-1,p},\quad\mathbf{P}\text{ a.s.}$$

Given a stopping time τ, we consider a stochastic interval

$$[[0,\tau]]=\begin{cases}[0,\tau(\omega)] & \text{if } \tau(\omega)<\infty,\\ [0,\infty) & \text{otherwise.}\end{cases}$$

DEFINITION 2.2. *Given a stopping time τ, an $\mathbb{H}^s_p(\mathbf{R}^d)$-valued $\mathbb{F}$-adapted function $\mathbf{u}(t)$ on $[0,\infty)$ is called an $\mathbb{H}^s_p$-solution of* (2.3) *in $[[0,\tau]]$ if it is strongly continuous in t with probability* 1,

$$\mathbf{u}(t\wedge\tau)=\mathbf{u}(t),\quad\operatorname{div}\mathbf{u}(t)=0,\quad\int_0^{t\wedge\tau}|\mathbf{u}(r)|^p_{s+1,p}\,dr<\infty\quad\forall t>0,\mathbf{P}\text{ a.s.},\tag{2.4}$$

and there exist two gradient vector fields, an $\mathbb{F}$-adapted $\mathcal{G}(\mathbb{H}^{s-1}_p(\mathbf{R}^d))$-valued process $\mathbf{G}(t)$ and an $\mathbb{F}$-adapted $\mathcal{G}(\mathbb{H}^{s-1}_p(\mathbf{R}^d,Y))$-valued process $\tilde{\mathbf{G}}(t)$, such that

$$\mathbf{G}=1_{[[0,\tau]]}\mathbf{G},\tilde{\mathbf{G}}=1_{[[0,\tau]]}\tilde{\mathbf{G}},\qquad dtd\mathbf{P}\text{ a.e.},$$

$$\int_0^{t\wedge\tau}[|\mathbf{G}(r)|^p_{s-1,p}+||\tilde{\mathbf{G}}(r)||^p_{s-1,p}]\,dr<\infty\quad\forall t>0,\ \mathbf{P}\text{ a.s.},$$

and the equality

$$\mathbf{u}(t\wedge\tau)=\mathbf{u_0}+\int_0^{t\wedge\tau}[\partial_i(a^{ij}(r)\partial_j\mathbf{u})+\mathbf{D}(\mathbf{u},r)-\mathbf{G}(r)]\,dr$$

$$+\int_0^{t\wedge\tau}[\sigma^k(r)\partial_k\mathbf{u}\,(r)+\mathbf{Q}(\mathbf{u},r)-\tilde{\mathbf{G}}(r)]\cdot dW\,(r) \tag{2.5}$$

holds in $\mathbb{H}_p^{s-1}(\mathbf{R}^d)$ *for every* $t>0$, $\mathbf{P}$ *a.s.*

If $\tau=\infty$, *we simply say* $\mathbf{u}$ *is an* $\mathbb{H}_p^s$*-solution of* (2.3).

The main result of the paper is the following statement.

THEOREM 2.3. *Let* $s\in(-\infty,\infty)$, $p\geq 2$. *Assume* A, A1(s,p)–A3(s,p) *are satisfied and* $\mathbf{E}|\mathbf{u}_0|^p_{s+1-2/p,p}<\infty$. *Then for each stopping time* τ, *the Cauchy problem* (1.1) *has a unique* $\mathbb{H}_p^s$*-solution in* $[[0,\tau]]$. *Moreover, the gradient processes in* (2.5) *are uniquely determined, and for each* $T>0$ *there is a constant* C *such that for each stopping time* $\bar\tau\leq T\wedge\tau$

$$\mathbf{E}\left[\sup_{r\leq\bar\tau}|\mathbf{u}(r)|^p_{s,p}+\int_0^{\bar\tau}(|\partial^2\mathbf{u}(r)|^p_{s-1,p}+|\mathbf{G}(r)|^p_{s-1,p}+||\tilde{\mathbf{G}}(r)||^p_{s,p})\,dr\right]$$

$$\leq C\mathbf{E}\left[|\mathbf{u}_0|^p_{s+1-2/p,p}+\int_0^{\bar\tau}(|\mathbf{D}(\mathbf{0},r)|^p_{s-1,p}+|\mathbf{Q}(\mathbf{0},r)|^p_{s,p})\,dr\right].$$

3. Some properties of function spaces.

3.1. Pointwise multipliers in $\mathbb{H}_p^s$. We will need the following statement about Hilbert space valued multipliers in $\mathbb{H}_p^s$, proved in [7].

LEMMA 3.1 (see Lemma 7 in [7]). (a) *Let* $a\in B^{|s|}(Y)$, $s\in(-\infty,\infty)$, $p\in(1,\infty)$. *Then there is a constant* N *so that*

$$||a\mathbf{u}||_{s,p}\leq N||a||_{B^{|s|}}|\mathbf{u}|_{s,p}$$

for all $\mathbf{u}\in\mathbb{H}_p^s$, *where* $a\mathbf{u}=(au^1,\dots,au^d)$.

(b) *Assume,* $p\in(1,\infty)$, $\kappa>0$, *and*

$$a\in\begin{cases}B^s(Y) & \text{if } s\geq 0,\\ B^{|s|+\kappa}(Y) & \text{if } s<0.\end{cases}$$

Let $\bar a_s=|a|_{B^s}$ *if* $s\geq 0$, *and* $\bar a_s=|a|_{B^{|s|+\kappa}}$ *if* $s<0$.

Then for every s *there exist constants* $s_0<s$ *and* N *such that*

$$||a\mathbf{u}||_{s,p}\leq N(||a||_\infty|\mathbf{u}|_{s,p}+\bar a_s|\mathbf{u}|_{s_0,p})$$

for all $\mathbf{u}\in\mathbb{H}_p^s$.

3.2. Solenoidal and potential projections of vector fields. First of all we decompose square integrable vector fields $\mathbf{v}\in\mathbb{L}_2(Y)=\mathbb{L}_2(\mathbf{R}^d,Y)$ (Y is a separable Hilbert space). Let

$$\mathcal{S}(\mathbb{L}_2(Y))=\{\mathbf{g}\in\mathbb{L}_2(Y):\operatorname{div}\mathbf{g}=0\},$$

where $\operatorname{div} \mathbf{g} = \partial_l g^l$. (All the component functions g^l are Y-valued.) Obviously, $\mathcal{S}(\mathbb{L}_2(Y))$ is a Hilbert subspace of $\mathbb{L}_2(Y)$, and

$$\mathbb{L}_2(Y) = \mathcal{G}(\mathbb{L}_2(Y)) \oplus \mathcal{S}(\mathbb{L}_2(Y)), \tag{3.1}$$

where $\mathcal{G}(\mathbb{L}_2(Y))$ is the orthogonal complement of $\mathcal{S}(\mathbb{L}_2(Y))$. Vector fields from $\mathcal{S}(\mathbb{L}_2(Y))$ are usually referred to as solenoidal or divergence free.

We will use a Riesz transform for the definition of solenoidal and gradient projections of a vector field. We set for $f \in L_p(\mathbf{R}^d, Y), 1 \le p < \infty$,

$$R_j(f)(x) = \lim_{\varepsilon \to 0} c_* \int_{|y| \ge \varepsilon} \frac{y_j}{|y|^{d+1}} f(x-y)\, dy, \quad j = 1, \dots, d,$$

with $c_* = G(\frac{n+1}{2})/\pi^{(n+1)/2}$ (G is the Gamma function). R_j is called a Riesz transform. According to [6, Chapter III, formula (8), p. 58],

$$(R_j \hat{f})(x) = -i \frac{\xi_j}{|\xi|} . \hat{f}, \tag{3.2}$$

where

$$\hat{f}(\xi) = \mathcal{F}(f) = (2\pi)^{-d/2} \int e^{-i(\xi, x)} f(x)\, dx.$$

Given a function $f \in L_p(\mathbf{R}^d, Y)$, we define a vector Riesz transform $Rf = (R_1 f, \dots, R_d f)$.

The following identity (see [6, Chapter III, Proposition 3, p. 59]) holds for each $u \in C_0^\infty(\mathbf{R}^d)$:

$$\partial_{jl}^2 u(x) = -R_l R_j \Delta u(x). \tag{3.3}$$

(The identity follows easily if we take the Fourier transform of (3.3) and use (3.2).)

LEMMA 3.2. *Let $\mathcal{G}$ be a projection of $\mathbb{L}_2(Y)$ onto $\mathcal{G}(\mathbb{L}_2(Y))$, and $\mathcal{S}$ be a projection of $\mathbb{L}_2(Y)$ onto $\mathcal{S}(\mathbb{L}_2(Y))$. Then*

$$\mathcal{G}(\mathbf{v}) = -RR_j v^j, \qquad \mathcal{S}(\mathbf{v}) = \mathbf{v} - \mathcal{G}(\mathbf{v}), \quad \mathbf{v} \in \mathbb{L}_2(Y).$$

Proof. By the Calderon–Zygmund theorem (see [6, Chapter II, Theorem 5, p. 46]) the Riesz transform is a bounded linear operator on L_2. Obviously, for each $\mathbf{v} \in \mathbb{L}_2(Y)$,

$$\mathbf{v} = -RR_j v^j + (\mathbf{v} + RR_j v^j).$$

Let us assume that $\mathbf{v}, \mathbf{g}, \partial \mathbf{g} \in \mathbb{L}_2(Y)$, $\operatorname{div} \mathbf{g} = 0$. Then

$$\mathcal{F}(\operatorname{div}(\mathbf{v} + RR_j v^j)) = i\xi_k \mathcal{F}(v^k) - i\xi_k \frac{\xi_k}{|\xi|} \frac{\xi_j}{|\xi|} \mathcal{F}(v^j) = 0,$$

and (by Parseval's equality)

$$\begin{aligned} \int \mathcal{G}(\mathbf{v}) \cdot \mathbf{g}\, dx &= \int \mathcal{F}(\mathcal{G}(\mathbf{v})) \cdot \bar{\mathcal{F}}(\mathbf{g})\, d\xi \\ &= \int \left(\frac{\xi}{|\xi|} \frac{\xi_k}{|\xi|} \mathcal{F}(v^k), \bar{\mathcal{F}}(\mathbf{g}) \right) d\xi = 0. \end{aligned}$$

So, the statement follows. □

Remark 3.3. If $f \in C_0^\infty(\mathbf{R}^d)$, it is known (see, e.g., [8]) that the classical solution to

$$\Delta u(x) = f(x), \quad x \in \mathbf{R}^d, \tag{3.4}$$

is given by the formula

$$u(x) = \int \Gamma(x-y) f(y)\, dy, \tag{3.5}$$

where

$$\Gamma(x-y) = \begin{cases} |x-y|^{2-d}/d(2-d)\omega_d, & d > 2, \\ \frac{1}{2\pi} \ln|x-y|, & d = 2, \end{cases}$$

and ω_d is the volume of the unit ball in $\mathbf{R}^d$. If $\mathbf{f} \in \mathbb{C}_0^\infty(Y)$, it is rather straightforward to show that

$$\mathcal{G}(\mathbf{f}) = \nabla \int \Gamma_{x_i}(x-y) f^i(y)\, dy = -RR_j f^j. \tag{3.6}$$

The functions $\mathcal{G}(\mathbf{v})$ and $\mathcal{S}(\mathbf{v})$ are usually referred to as the potential and the solenoidal, respectively, projections, of the vector field $\mathbf{v}$.

COROLLARY 3.4. *For any* $\mathbf{v}, \mathbf{u} \in \mathbb{L}_2(Y)$,

$$\int \mathcal{G}(\mathbf{u}) \cdot \mathbf{v}\, dx = \int \mathbf{u} \cdot \mathcal{G}(\mathbf{v})\, dx. \tag{3.7}$$

Proof. Indeed, because of orthogonality, both integrals are equal to $\int \mathcal{G}(\mathbf{u}) \cdot \mathcal{G}(\mathbf{v})\, dx$. □

Remark 3.5. Let $p \in (1, \infty)$. Note that, by Calderón–Zygmund's inequality [6, Chapter II, Theorem 5, p. 46], the Riesz transform is bounded on $\mathbb{L}_p(Y)$. Therefore, the function $\mathcal{G}(\mathbf{f}) = \nabla \int \Gamma_{x_i}(x-y) f^i(y)\, dy = -RR_i f^i$ is defined for all $\mathbf{f} \in \mathbb{L}_p(Y)$, and there is a constant C such that

$$||\mathcal{G}(\mathbf{f})||_p \le C||\mathbf{f}||_p, \quad \mathbf{f} \in \mathbb{L}_p(Y). \tag{3.8}$$

LEMMA 3.6. *Let* $1 < p < \infty$. *Then for each* $\mathbf{v} \in \mathbb{C}_0^\infty(Y)$ *we have* $\mathcal{G}(\mathbf{v}), \mathcal{S}(\mathbf{v}) \in \cap_s \mathbb{H}_p^s(Y)$ *and*

$$\begin{aligned} (1-\Delta)^{s/2}\mathcal{G}(\mathbf{v}) &= \mathcal{G}((1-\Delta)^{s/2}\mathbf{v}), \\ (1-\Delta)^{s/2}\mathcal{S}(\mathbf{v}) &= \mathcal{S}((1-\Delta)^{s/2}\mathbf{v}). \end{aligned} \tag{3.9}$$

Moreover, there is a constant C *so that for all* $\mathbf{v} \in \mathbb{C}_0^\infty(Y)$

$$||\mathcal{G}(\mathbf{v})||_{s,p} \le C||\mathbf{v}||_{s,p}, ||\mathcal{S}(\mathbf{v})||_{s,p} \le C||\mathbf{v}||_{s,p}$$

for any $s \in (-\infty, \infty)$ *and* $\mathcal{G}, \mathcal{S}$ *can be extended by continuity to all* $\mathbb{H}_p^s(Y)$, $s \in (-\infty, \infty)$.

Also, for each $\mathbf{v} \in \mathbb{H}_p^s(Y)$,

$$\mathcal{G}(\mathbf{v}) = (1-\Delta)^{-s/2}\mathcal{G}((1-\Delta)^{s/2}\mathbf{v}).$$

Proof. For $\mathbf{v} \in \mathbb{C}_0^\infty(Y)$, we have

$$\mathcal{F}((1-\Delta)^{s/2}\mathcal{G}(\mathbf{v})) = (1+|\xi|^2)^{s/2}\frac{\xi}{|\xi|}\frac{\xi_k}{|\xi|}\mathcal{F}(v^k) \in \mathbb{L}_2(Y)$$

for each s, and $\mathcal{F}((1-\Delta)^{s/2}\mathcal{G}(\mathbf{v})) = \mathcal{F}(\mathcal{G}((1-\Delta)^{s/2}\mathbf{v}))$. Therefore, (3.9) holds. According to Remark 3.5 and (3.8),

$$||\mathcal{G}(\mathbf{v})||_{s,p} = ||\mathcal{G}((1-\Delta)^{s/2}\mathbf{v})||_p \le C||(1-\Delta)^{s/2}\mathbf{v}||_p = C||\mathbf{v}||_{s,p}.$$

Since $\mathcal{S}(\mathbf{v}) = \mathbf{v} - \mathcal{G}(\mathbf{v})$, we can immediately obtain the extensions to $\mathbb{H}_p^s(Y)$. ∎

The following statement is the direct consequence of Lemma 3.6.

LEMMA 3.7. *Suppose $p \in (1,\infty)$ and $s \in (-\infty,\infty)$. Then the space $\mathbb{H}_p^s(Y)$ can be decomposed into the direct sum*

$$\mathbb{H}_p^s(Y) = \mathcal{G}(\mathbb{H}_p^s(Y)) \oplus \mathcal{S}(\mathbb{H}_p^s(Y)).$$

Moreover, if $1/q + 1/p = 1$, $\mathbf{f} \in \mathcal{G}(\mathbb{H}_p^s(Y))$, $\mathbf{g} \in \mathcal{S}(\mathbb{H}_q^{-s}(Y))$, then

$$\langle \mathbf{f}, \mathbf{g}\rangle_{\mathbb{H}_p^s(Y),\mathbb{H}_q^{-s}(Y)} = 0. \tag{3.10}$$

Also,

$$\mathcal{S}(\mathbb{H}_p^s(Y)) = \{\mathbf{v} \in \mathbb{H}_p^s(Y) : \operatorname{div}\mathbf{v} = \mathbf{0}\}, \tag{3.11}$$

$$\mathcal{G}(\mathbb{H}_p^s(Y)) = \{\mathbf{v} \in \mathbb{H}_p^s(Y) : \langle \mathbf{v}, \mathbf{g}\rangle_{\mathbb{H}_p^s(Y),\mathbb{H}_q^{-s}(Y)} = 0 \quad \forall \mathbf{g} \in \mathcal{S}(\mathbb{H}_{\mathbf{q}}^{-\mathbf{s}}(\mathbf{Y}))\}. \tag{3.12}$$

Proof. Let $1/q + 1/p = 1$, $\tilde{\mathbf{f}} \in \mathbb{H}_p^s(Y)$, $\tilde{\mathbf{g}} \in \mathbb{H}_q^{-s}(Y)$ and $\mathbf{f} = \mathcal{G}(\tilde{\mathbf{f}})$, $\mathbf{g} = \mathcal{S}(\tilde{\mathbf{g}})$. By Lemma 3.6, there are some sequences $\mathbf{f}_n \in \mathbb{C}_0^\infty$ and $\mathbf{g}_n \in \mathbb{C}_0^\infty$ such that

$$|\mathbf{f}_n - \tilde{\mathbf{f}}|_{s,p} + |\mathbf{g}_n - \tilde{\mathbf{g}}|_{-s,q} \to 0.$$

Then

$$|\mathcal{G}(\mathbf{f}_n) - \mathcal{G}(\tilde{\mathbf{f}})|_{s,p} + |\mathcal{S}(\mathbf{g}_n) - \mathcal{S}(\tilde{\mathbf{g}})|_{-s,q} \to 0,$$

and, by Lemmas 3.6 and 3.2,

$$\begin{aligned}\langle \mathbf{f}, \mathbf{g}\rangle_{\mathbb{H}_p^s(Y),\mathbb{H}_q^{-s}(Y)} &= \int (1-\Delta)^{s/2}\mathcal{G}(\tilde{\mathbf{f}})(1-\Delta)^{-s/2}\mathcal{S}(\tilde{\mathbf{g}})\,dx \\ &= \lim_n \int (1-\Delta)^{s/2}\mathcal{G}(\mathbf{f}_n)(1-\Delta)^{-s/2}\mathcal{S}(\mathbf{g}_n)\,dx \\ &= \lim_n \int \mathcal{G}((1-\Delta)^{s/2}\mathbf{f}_n)\mathcal{S}((1-\Delta)^{-s/2}\mathbf{g}_n)\,dx = 0.\end{aligned}$$

So (3.10) holds.

If $\mathbf{f} \in \mathbb{H}_p^s(Y)$, we have, obviously, $\mathbf{f} = \mathcal{G}(\mathbf{f}) + [\mathbf{f} - \mathcal{G}(\mathbf{f})] = \mathcal{G}(\mathbf{f}) + \mathcal{S}(\mathbf{f})$. Now we prove that

$$\mathcal{G}(\mathbb{H}_p^s(Y)) \cap \mathcal{S}(\mathbb{H}_p^s(Y)) = \{\mathbf{0}\}.$$

Suppose $\mathbf{f} \in \mathcal{G}(\mathbb{H}^s_p(Y)) \cap \mathcal{S}(\mathbb{H}^s_p(Y))$. Then for each $\mathbf{v} \in \mathbb{C}^\infty_0(Y)$ we have by Lemma 3.6 and (3.10) that

$$\int (1-\Delta)^{s/2}\mathbf{f}\cdot(1-\Delta)^{-s/2}\mathcal{S}(\mathbf{v})\,dx = \int (1-\Delta)^{s/2}\mathbf{f}\cdot\mathcal{S}((1-\Delta)^{-s/2}\mathbf{v})\,dx$$

$$= \langle \mathbf{f}, \mathcal{S}(\mathbf{v})\rangle_{\mathbb{H}^s_p(Y),\mathbb{H}^{-s}_q(Y)} = 0.$$

Also,

$$\int (1-\Delta)^{s/2}\mathbf{f}\cdot(1-\Delta)^{-s/2}\mathcal{G}(\mathbf{v})\,dx = \int (1-\Delta)^{s/2}\mathbf{f}\cdot\mathcal{G}((1-\Delta)^{-s/2}\mathbf{v})\,dx$$

$$= \langle \mathbf{f}, \mathcal{G}(\mathbf{v})\rangle_{\mathbb{H}^s_p(Y),\mathbb{H}^{-s}_q(Y)} = 0.$$

Therefore, $\langle \mathbf{f}, \mathbf{v}\rangle_{\mathbb{H}^s_p(Y),\mathbb{H}^{-s}_q(Y)} = \langle \mathbf{f}, \mathcal{G}(\mathbf{v}) + \mathcal{S}(\mathbf{v})\rangle_{\mathbb{H}^s_p(Y),\mathbb{H}^{-s}_q(Y)} = 0$. Thus, $\mathbf{f} = \mathbf{0}$.

Now, we prove (3.11). Let $\mathbf{v} \in \mathbb{H}^s_p(Y)$, $\operatorname{div}\mathbf{v} = 0$. Let $\varphi \in C^\infty_0(\mathbf{R}^d)$ be a scalar nonnegative function so that $\int \varphi\,dx = 1$. For $\varepsilon > 0$, write $\varphi_\varepsilon(x) = \varepsilon^{-d}\varphi(x/\varepsilon)$. Set

$$\mathbf{v}_k(x) = (1-\Delta)^{-s/2}\int \varphi_{1/k}(x-y)(1-\Delta)^{s/2}\mathbf{v}(y)$$

$$= \int (1-\Delta)^{-s/2}\varphi_{1/k}(x-y)(1-\Delta)^{s/2}\mathbf{v}(y)\,dy$$

$$= \int (1-\Delta)^{-s/2}\varphi_{1/k}(y)(1-\Delta)^{s/2}\mathbf{v}(x-y)\,dy.$$

Obviously, $\mathbf{v}_k \in \mathbb{H}^{s+1}_p(Y)$, $\operatorname{div}\mathbf{v}_k = 0$, and $|\mathbf{v}_k - \mathbf{v}|_{n,p} \to 0$ as $k \to \infty$. By Corollary 3.4, $\mathcal{G}(\mathbf{v}_k) = \mathbf{0}$. Therefore, $\mathcal{G}(\mathbf{v}) = \mathbf{0}$ and $\mathbf{v} = \mathcal{S}(\mathbf{v})$.

Let $\mathbf{v} \in \mathbb{H}^s_p(Y)$ and $\langle \mathbf{v}, \mathbf{g}\rangle_{\mathbb{H}^s_p(Y),\mathbb{H}^{-s}_q(Y)} = 0$ for all $\mathbf{g} \in \mathcal{S}(\mathbb{H}^{-\mathbf{s}}_{\mathbf{q}}(\mathbf{Y}))$. Then for any $\mathbf{h} \in \mathbb{H}^{-\mathbf{s}}_{\mathbf{q}}(\mathbf{Y})$

$$0 = \langle \mathbf{v}, \mathcal{S}(\mathbf{h})\rangle_{\mathbb{H}^s_p(Y),\mathbb{H}^{-s}_q(Y)} = \langle \mathcal{S}(\mathbf{v}), \mathbf{h}\rangle_{\mathbb{H}^s_p(Y),\mathbb{H}^{-s}_q(Y)}.$$

Therefore, $\mathcal{S}(\mathbf{v}) = \mathbf{0}$, i.e., $\mathbf{v} = \mathcal{G}(\mathbf{v})$, and equality (3.12) follows. □

LEMMA 3.8. *Assume* $\mathbf{v} \in \mathbb{H}^{s+1}_p(Y)$, $p \in (1,\infty)$. *Then*

$$\mathcal{G}(\partial_l \mathbf{v}) = \partial_l \mathcal{G}(\mathbf{v}) = -(1-\Delta)^{-s/2}\mathbf{R}R_l((1-\Delta)^{s/2}\operatorname{div}\mathbf{v}). \tag{3.13}$$

There is a constant C *such that for all* $\mathbf{v} \in \mathbb{H}^{s+1}_p(Y)$

$$||\partial\mathcal{G}(\mathbf{v})||_{s,p} \le C||\operatorname{div}\mathbf{v}||_{s,p},$$

and for all $\mathbf{v} \in \mathbb{H}^s_p(Y)$

$$||\mathcal{G}(\mathbf{v})||_{s,p} \le C||\operatorname{div}\mathbf{v}||_{s-1,p} + ||\mathbf{v}||_{s-1,p}.$$

Proof. By Lemma 3.6 and Remark 3.5, we have for $\tilde{\mathbf{v}} = \partial_l \mathbf{v}$ that

$$\mathcal{G}(\tilde{\mathbf{v}}) = -(1-\Delta)^{-s/2}RR_k((1-\Delta)^{s/2}\tilde{v}^k)$$
$$= -(1-\Delta)^{-s/2}RR_k((1-\Delta)^{s/2}\partial_l v^k).$$

Notice that

$$\begin{aligned} &-(1-\Delta)^{-s/2}RR_k((1-\Delta)^{s/2}\partial_l v^k) \\ &= -(1-\Delta)^{-s/2}RR_l((1-\Delta)^{s/2}\operatorname{div}\mathbf{v}) \\ &= -\partial_l(1-\Delta)^{-s/2}RR_k((1-\Delta)^{s/2}v^k). \end{aligned} \tag{3.14}$$

Indeed, if $\mathbf{v}\in\mathbb{C}_0^\infty(Y)$, taking the Fourier transform of each term in (3.14), we have

$$\frac{\xi}{|\xi|}\frac{\xi_k}{|\xi|}i\xi_l\mathcal{F}(v^k)=\frac{\xi}{|\xi|}\frac{\xi_l}{|\xi|}i\xi_k\mathcal{F}(v^k)=i\xi_l\frac{\xi}{|\xi|}\frac{\xi k}{|\xi|}\mathcal{F}(v^k).$$

So, the first part of the statement follows, and we have, obviously, the first inequality. Since s is arbitrary, we have

$$||\mathcal{G}(\mathbf{v})||_{s,p}\le C\bigg(\sum_l||\partial_l\mathcal{G}(\mathbf{v})||_{s-1,p}+||\mathbf{v}||_{s-1,p}\bigg)\le C(||\operatorname{div}\mathbf{v}||_{s-1,p}+||\mathbf{v}||_{s-1,p})$$

by (3.13) and Remark 3.5. □

Later we will need L_p-estimates of the function $\mathcal{G}(\mathbf{h})$, where $\mathbf{h}=c^j\partial_j\mathbf{v}$.

LEMMA 3.9. *Let* $\mathbf{h}=c^j(x)\partial_j\mathbf{v}(x)$, *where* $c=(c^j)$ *is a measurable d-vector of Hilbert space* Y*-valued functions,* $\mathbf{v}\in\mathbb{H}_p^{s+1}$, $\operatorname{div}\mathbf{v}=0$, $\varepsilon\in(0,1)$. *Assume*

$$\begin{aligned} ||c||_{B^{|s|}}&<\infty \text{ if } s\ge 1,\\ ||c||_{B^1}&<\infty \text{ if } s\in(-1,1),\\ ||c||_{B^{-s+\varepsilon}}&<\infty \text{ if } s\le -1. \end{aligned}$$

Then

$$||\mathcal{G}(\mathbf{h})||_{s,p}\le\begin{cases} C(||\partial_l c^j\partial_j\mathbf{v}||_{s-1,p}+||c^j\partial_j\mathbf{v}||_{s-1,p}) & \text{if } s>0,\\ C(||\partial_l c^j v^l||_{s,p}+||\operatorname{div}\mathrm{c}\,\mathbf{v}||_{s,p}] & \text{if } s\le 0. \end{cases}$$

Proof. Let $s>0$. Then, by inequality (3.13) of Lemma 3.8,

$$||\mathcal{G}(\mathbf{h})||_{s,p}\le C(||\operatorname{div}\mathbf{h}||_{s-1,p}+||\mathbf{h}||_{s-1,p})$$

and $\operatorname{div}\mathbf{h}=\partial_l c^j\partial_j v^l$. Let $s\le 0$. Since

$$c^j\partial_j\mathbf{v}=\partial_j(c^j\mathbf{v})-\partial_j\,c^j\mathbf{v},$$

it follows by Lemma 3.8 and Remark 3.5 that

$$\begin{aligned} ||\mathcal{G}(\mathbf{h})||_{s,p}&\le||\partial_j\mathcal{G}(c^j\mathbf{v})||_{s,p}+||\mathcal{G}(\partial_j c^j\mathbf{v})||_{s,p}\\ &\le C(||\partial_l c^j v^l||_{s,p}+||\partial_j c^j\,\mathbf{v}||_{s,p}), \end{aligned}$$

and the second inequality follows. □

Also we will need L_p-estimates of the function $\mathcal{G}(\mathbf{h})$, where $\mathbf{h}=\partial_i(c^{ij}(x)\partial_j\mathbf{v})$.

COROLLARY 3.10. *Let* $\mathbf{h}=\partial_i(c^{ij}(x)\partial_j\mathbf{v})$, *where* $c=(c^{ij})$ *is a measurable function,* $\mathbf{v}\in\mathbb{H}_p^{s+1}$, $\operatorname{div}\mathbf{v}=0$, $\varepsilon\in(0,1)$. *Assume*

$$\begin{aligned} |c|_{B^{|s|}}&<\infty \text{ if } s\ge 1,\\ |c|_{B^1}&<\infty \text{ if } s\in(-1,1),\\ |c|_{B^{-s+\varepsilon}}&<\infty \text{ if } s\le -1. \end{aligned}$$

Then

$$|\mathcal{G}(\mathbf{h})|_{s-1,p} \le \begin{cases} C(|\partial_l c^{ij}\partial_j v^l|_{s-1,p} + |c^{ij}\partial_j v^l|_{s-1,p}) & \text{if } s > 0, \\ C(|\partial_l c^{ij} v^j|_{s,p} + |\partial_j c^{ij}\,\mathbf{v}|_{s,p}] & \text{if } s \le 0. \end{cases}$$

Proof. Indeed,

$$|\mathcal{G}(\mathbf{h})|_{s-1,p} \le C\sum_i |\mathcal{G}\left(c^{ij}(x)\partial_j \mathbf{v}\right)|_{s,p},$$

and the inequality follows by Lemma 3.9. □

4. Stochastic Stokes equation. We rewrite (2.3) in an equivalent form:

$$\begin{aligned} \partial_t \mathbf{u}(t,x) &= \mathcal{S}(\partial_i(a^{ij}(t,x)\partial_j \mathbf{u}) + \mathbf{D}(\mathbf{u},t,x)) \\ &\quad \mathcal{S}(\sigma^k(t,x)\partial_k \mathbf{u}(t,x) + \mathbf{Q}(\mathbf{u},t,x))\cdot \dot{W}, \\ \mathbf{u}(0,x) &= \mathbf{u}_0(x). \end{aligned} \tag{4.1}$$

We use the following equivalent definition of an $\mathbb{H}^s_p$-solution of (2.3) (or (4.1)).

DEFINITION 4.1. *Given a stopping time τ, an $\mathbb{H}^s_p(\mathbf{R}^d)$-valued $\mathbb{F}$-adapted function $\mathbf{u}(t)$ on $[0,\infty)$ is called an $\mathbb{H}^s_p$-solution of* (2.3) (*or* (4.1)) *in $[[0,\tau]]$ if it is strongly continuous in t with probability* 1,

$$\mathbf{u}(t\wedge\tau) = \mathbf{u}(t), \qquad \int_0^{t\wedge\tau} |\mathbf{u}(r)|^p_{s+1,p}\,dr < \infty \quad \forall t > 0, \mathbf{P}\ \textit{a.s.}, \tag{4.2}$$

and the equality

$$\begin{aligned} \mathbf{u}(t\wedge\tau) = \mathbf{u_0} &+ \textstyle\int_0^{t\wedge\tau} \mathcal{S}(\partial_i(a^{ij}(r)\partial_j \mathbf{u}) + \mathbf{D}(\mathbf{u},r))dr \\ &+ \textstyle\int_0^{t\wedge\tau} \mathcal{S}(\sigma^k(r)\partial_k \mathbf{u}\,(r) + \mathbf{Q}(\mathbf{u},r))\cdot dW\,(r) \end{aligned} \tag{4.3}$$

holds in $\mathbb{H}^{s-1}_p(\mathbf{R}^d)$ for every $t > 0$, $\mathbf{P}$ a.s.

If $\tau = \infty$, we simply say $\mathbf{u}$ is an $\mathbb{H}^s_p$-solution of (2.3).

It is readily checked that all the integrals in (4.3) are well defined. For example, let us consider the stochastic integral. By (4.2),

$$\int_0^{t\wedge\tau} |\mathbf{u}(r)|^p_{s+1,p}\,dr < \infty \quad \forall t > 0, \mathbf{P}\text{ a.s.} \tag{4.4}$$

Since ∂_i is a bounded operator from $\mathbb{H}^m_p$ into $\mathbb{H}^{m-1}_p$ (see [5]), by Lemma 5.2 in [2] and assumption A1(s,p), we have $||\mathcal{S}(\sigma^k(r)\partial_k \mathbf{u}(r))||_{s,p} \le C||\mathbf{u}(r)||_{s+1,p}$. By assumptions A2$(s,p)$ and A3(s,p),

$$\int_0^{t\wedge\tau} ||\mathcal{S}(\mathbf{Q}(\mathbf{u},r))||^p_{s,p}\,dr \le C\int_0^{t\wedge\tau} (||\mathbf{Q}(\mathbf{0},r)||^p_{s,p} + |\mathbf{u}\,(r)\,|^p_{s+1,p})dr.$$

Thus, the integral is defined according to (2.2).

Remark 4.2. It is not difficult to show that (4.3) can be replaced by the equality

$$\begin{aligned} &\left\langle u^l(t\wedge\tau), \phi^l\right\rangle_s = \left\langle u^l_0, \phi^l\right\rangle_s + \textstyle\int_0^{t\wedge\tau} -\left\langle \mathcal{S}(a^{ij}(r)\partial_i u^l), \partial_j \phi^l\right\rangle_s \\ &+ \langle \Lambda^{-1} D^l(\mathbf{u},r), \Lambda\phi^l\rangle_s dr + \textstyle\int_0^{t\wedge\tau} \langle \mathcal{S}(\sigma^k(r)\partial_k u^l + Q^l(\mathbf{u},r)), \phi^l\rangle_{s,Y}\cdot dW\,(r) \\ &\qquad \forall t > 0, \mathbf{P}\text{ a.s.}, \end{aligned} \tag{4.5}$$

which holds for all $\boldsymbol{\phi} = (\phi^l)_{1\leq l\leq d}$ such that $\phi^l \in C_0^\infty, l = 1, \dots, d$.

Indeed, owing to (2.5), we have

(4.6)
$$\left(u^l(t\wedge\tau),\phi^l\right)_{s-1} = \left(u_0^l,\phi^l\right)_{s-1} + \int_0^{t\wedge\tau}\left(\mathcal{S}(\partial_j(a^{ij}(r)\partial_i u^l) + D^l(\mathbf{u},r)),\phi^l\right)_{s-1} dr$$

$$+\int_0^{t\wedge\tau}\left(\mathcal{S}(\sigma^k(r)\partial_k u^l + Q^l(\mathbf{u},r)),\phi^l\right)_{s-1,Y}\cdot dW(r) \quad \forall t>0, \mathbf{P}\text{ a.s.}$$

On the other hand, since $\mathbf{u}\in\mathbb{H}_p^s$, $\mathbf{u}_0\in\mathbb{H}_p^{s+1-2/p}$, and for almost all r, $\sigma^k(r)\partial_k\mathbf{u}(r)+Q^l(\mathbf{u},r)\in\mathbb{H}_p^s$, $\mathbf{P}$ a.s., we have that

$$\left(u^l(t),\phi^l\right)_{s-1} = \left(u^l(t),\phi^l\right)_s, \left(u_0^l,\phi^l\right)_{s-1} = \left(u_0^l,\phi^l\right)_{s+1-2/p},$$

and for almost all s,

$$\left(\sigma^k(r)\partial_k u^l(r) + Q^l(\mathbf{u},r),\phi^l\right)_{s-1,Y} = \left(\sigma^k(r)\partial_k u^l(r) + Q^l(\mathbf{u},r),\phi^l\right)_{s,Y}$$

$\mathbf{P}$ a.s. It is readily checked that $dr\times d\mathbf{P}$ a.e.

$$\left(\partial_j(a^{ij}(r)\partial_i u^l),\phi^l\right)_{s-1} = \left(\partial_j(a^{ij}(r)\partial_i u^l),\phi^l\right)_{s-1}$$

$$= -\left(\Lambda^s(a^{ij}(r)\partial_i u^l),\Lambda^{-s}\partial_j\phi^l\right)_0 = -\left((a^{ij}(r)\partial_i u^l),\partial_j\phi^l\right)_s.$$

Note that to prove the first equality, one should first establish it for smooth functions and then prove it in the general case by approximations. Thus, (4.6) implies (4.5). Now by reversing the order of our arguments, one could easily show that (4.3) follows from (4.5).

The main existence theorem will be proved in several steps. We begin with a simple particular case.

THEOREM 4.3 (cf. Theorem 4.10 in [2]). *Assume* A, A1(s,p)–A3(s,p). *Suppose that* $\mathbf{D}$ *and* $\mathbf{Q}$ *are independent of* $\mathbf{u}$, a^{ij} *and* σ^k *are independent of* x, $\mathbf{u}_0 = \mathbf{0}$, *and* $\operatorname{div}\mathbf{D}(t) = \operatorname{div}\mathbf{Q}(t) = 0$.

Then for each stopping time τ *there is a unique* $\mathbb{H}_p^s$*-solution* $\mathbf{u}$ *of* (4.1) *in* $[[0,\tau]]$. *Moreover,*

(i) *for each stopping time* $\bar\tau\leq\tau$,

$$\mathbf{E}\int_0^{\bar\tau}|\partial^2\mathbf{u}(r)|_{s-1,p}^p\,dr \leq N\mathbf{E}\int_0^{\bar\tau}(|\mathbf{D}(r)|_{s-1,p}^p + ||\mathbf{Q}(r)||_{s,p}^p)\,dr, \tag{4.7}$$

where $N = N(d,p,\delta,K)$ *does not depend on* τ;

(ii) *for each finite* T *and each stopping time* $\bar\tau\leq T\wedge\tau$,

$$\mathbf{E}\sup_{r\leq\bar\tau}|\mathbf{u}(r)|_{s,p}^p \leq e^T C\mathbf{E}\int_0^{\bar\tau}(|\mathbf{D}(r)|_{s-1,p}^p + ||\mathbf{Q}(r)||_{s,p}^p)\,dr, \tag{4.8}$$

where $C = C(d,p,\delta,K)$ *does not depend on* T *and* $\bar\tau,\tau$.

Proof. Consider the system

$$\begin{aligned}\partial_t\mathbf{v}(t,x) &= \partial_i(a^{ij}(t)\partial_j\mathbf{v}) + \mathbf{D}(t,x)\\ &\quad + [\sigma^k(t)\partial_k\mathbf{v}(t,x) + \mathbf{Q}(t,x)]\cdot\dot W,\\ \mathbf{v}(0,x) &= \mathbf{0}.\end{aligned} \tag{4.9}$$

According to [7], there is a unique $\mathbb{H}^s_p$-valued continuous $\mathbb{F}$-adapted solution to (4.9) in $[[0,\tau]]$ such that the estimates (4.7) and (4.8) hold. Then $\mathbf{u}(t)=\mathcal{S}(\mathbf{v}(t))$ is $\mathbb{H}^s_p$-valued continuous $\mathbb{F}$-adapted and (4.7) holds. According to our assumptions, $\mathbf{u}(t)$ satisfies the same equation (4.9). Therefore, $\mathbf{u}(t)=\mathcal{S}(\mathbf{v}(t))=\mathbf{v}(t)$, and the statement follows. □

To prove the general Theorem 2.3 we will rely on the two fundamental techniques: partition of unity and the method of continuity. The same technique was used in [2] for scalar equations.

The next step is to derive a priori L_p-estimates for a solution of (4.1).

LEMMA 4.4. *Assume* A, A1(s,p)–A3(s,p). *Suppose that* $\mathbf{u}$ *is an* $\mathbb{H}^s_p$*-solution of* (2.3) *in* $[[0,\tau]]$ *with* $\mathbf{u}_0=\mathbf{0}$.

Then for each T *there is a constant* $C=C(d,p,\delta,K,T)$ *such that for each stopping time* $\bar\tau\le T\wedge\tau$,

$$\begin{aligned}&\mathbf{E}[\sup_{r\le\bar\tau}|\mathbf{u}(r)|^p_{s,p}+\textstyle\int_0^{\bar\tau}|\partial^2\mathbf{u}(r)|^p_{s-1,p}\,dr]\\&\le C\mathbf{E}\textstyle\int_0^{\bar\tau}(|\mathbf{D}(\mathbf{0},r)|^p_{s-1,p}+|\mathbf{Q}(\mathbf{0},r)|^p_{s,p})\,dr.\end{aligned}\tag{4.10}$$

Proof. In order to use Theorem 4.3 we start with a standard partition of unity. Let $\psi\in C_0^\infty(\mathbf{R})$ be $[0,1]$-valued and such that $\psi(s)=1$, if $|s|\le 5/8$, and $\psi(s)=0$, if $|s|>6/8$. For an arbitrary but fixed $\kappa>0$ there we choose m such that $\kappa<2^{-m}$. Consider a grid in $\mathbf{R}^d$ consisting of $x_k=k2^{-m}$, $k=(k_1,\dots,k_d)\in\mathbf{Z}^d$, where $\mathbf{Z}$ is the set of all integers. Given $k\in\mathbf{Z}^d$, we define a function on $\mathbf{R}^d$ as

$$\bar\eta_k(x)=\prod_{l=1}^d\psi((x^l-x^l_k)2^m).$$

Notice that $0\le\bar\eta_k\le 1,\bar\eta_k=1$ in the cube $v_k=\{x:|x^l-x^l_k|\le(5/8)2^{-m},l=1,\dots,d\}$, and $\bar\eta_k=0$ outside the cube $V_k=\{x:|x^l-x^l_k|\le(6/8)2^{-m},l=1,\dots,d\}$. Obviously,

1. $\cup_k v_k=\mathbf{R}^d$ and

$$1\le\sum_k 1_{V_k}\le 2^d;$$

2. for all multi-indices γ

$$|\partial^\gamma\bar\eta_k|\le N(d,|\gamma|)2^{m|\gamma|}<N(d)\kappa^{-|\gamma|}.$$

Denote

$$\eta_k(x)=\bar\eta_k(x)\Big(\sum_k\bar\eta_k(x)\Big)^{-1},\qquad k=1,\dots.$$

Obviously, $\sum_k\eta_k=1$ in $\mathbf{R}^d$, and for all k and multi-indices μ,

$$|\partial^\mu\eta_k|\le N(d,|\mu|)\kappa^{-|\mu|},$$

and for each $p\ge 1,\mu$,

$$\sum_k\eta_k(x)^p\le N(p,d),\qquad\sum_k|\partial^\mu\eta_k|^p\le N(p,d,|\mu|)\kappa^{-p|\mu|}.\tag{4.11}$$

So, by Lemma 6.7 in [2], for any n there exist constants $c = c(d,p,\kappa)$, $C = C(d,p,\kappa)$ such that for all $\mathbf{f} \in \mathbb{H}^n_p, \mathbf{g} \in \mathbb{H}^n_p(Y)$

$$c|\mathbf{f}|^p_{n,p} \le \sum_k |\eta_k \mathbf{f}|^p_{n,p} \le C|\mathbf{f}|^p_{n,p}, \tag{4.12}$$

$$c||\mathbf{g}||^p_{n,p} \le \sum_k ||\eta_k \mathbf{g}||^p_{n,p} \le C||\mathbf{g}||^p_{n,p}.$$

Multiplying (4.1) by η_k and taking a solenoidal projection, we have

$$\begin{aligned} \partial_t \mathcal{S}(\eta_k \mathbf{u}) = {} & \partial_i(a^{ij}(t,x_k)\partial_j \mathcal{S}(\eta_k \mathbf{u})) + \mathcal{S}(\mathbf{D}_k(\mathbf{u},t,x)) \\ & + [\sigma^i(t,x_k)\partial_i \mathcal{S}(\eta_k \mathbf{u}) + \mathcal{S}(\mathbf{Q}_k(\mathbf{u},t,x))] \cdot \dot{W}, \end{aligned} \tag{4.13}$$

where

$$\begin{aligned} \mathbf{D}_k(\mathbf{u},t,x) = {} & \eta_k[\mathcal{S}(\mathbf{D}(\mathbf{u},t)) + \mathcal{S}(\partial_i(a^{ij}(t) - a^{ij}(t,x_k))\partial_j \mathbf{u}(t))] \\ & - \partial_i(a^{ij}(t,x_k)\partial_j \eta_k \mathbf{u}) - a^{ij}(t,x_k)\partial_i \eta_k \partial_j \mathbf{u}, \\ \mathbf{Q}_k(\mathbf{u},t,x) = {} & \eta_k[\mathcal{S}(\mathbf{Q}(\mathbf{u},t)) + \mathcal{S}((\sigma^i(t) - \sigma^i(t,x_k))\partial_i \mathbf{u}(t))] \\ & - \sigma^i(t,x_k)\partial_i \eta_k \mathbf{u}. \end{aligned}$$

We have

$$\sum_k |\eta_k \mathcal{G}(\partial_i[(a^{ij}(t) - a^{ij}(t,x_k))\partial_j \mathbf{u}(t)])|^p_{s-1,p}$$

$$\le C|\mathcal{G}(\partial_i[(a^{ij}(t) - a^{ij}(t,x_k))\partial_j \mathbf{u}(t)])|^p_{s-1,p},$$

and

$$\sum_k |\eta_k \mathcal{G}((\sigma^i(t) - \sigma^i(t,x_k))\partial_i \mathbf{u}(t))|^p_{s,p}$$

$$\le C|\mathcal{G}((\sigma^i(t) - \sigma^i(t,x_k))\partial_i \mathbf{u}(t))|^p_{s,p}.$$

Also,

$$\begin{aligned} & \sum_k |\eta_k \partial_i(a^{ij}(t) - a^{ij}(t,x_k)\partial_j \mathbf{u}(t))|^p_{s-1,p} \\ & \le 2^{p-1} \sum_k |\partial_i \eta_k (a^{ij}(t) - a^{ij}(t,x_k))\partial_j \mathbf{u}(t)|^p_{s-1,p} \\ & + 2^{p-1} \sum_k |\partial_i[\eta_k(a^{ij}(t) - a^{ij}(t,x_k))\tilde{\eta}_k \partial_j \mathbf{u}(t)]|^p_{s-1,p}, \end{aligned}$$

where $\tilde{\eta}_k(x) = \bar{\eta}_k(5x/6)$. (Notice that $\tilde{\eta}_k(x) = 1$ in V_k and $\tilde{\eta}_k(x) = 0$ if there is l such that $|x^l - x^l_k| > 0.9 \cdot 2^{-m}$.) According to Lemma 3.1, there is a constant C and $s_0 < s$ such that

$$\begin{aligned} & \sum_k |\partial_i[\eta_k(a^{ij}(t) - a^{ij}(t,x_k))\tilde{\eta}_k \partial_j \mathbf{u}(t)]|^p_{s-1,p} \\ & \le \sum_k |\eta_k(a^{ij}(t) - a^{ij}(t,x_k))\tilde{\eta}_k \partial_j \mathbf{u}(t)|^p_{s,p} \\ & \le C \sum_k [\sup_{x,k} |\tilde{\eta}_k(a^{ij}(t) - a^{ij}(t,x_k))|^p |\partial \mathbf{u}(t)|^p_{s,p} + |\eta_k \partial_j \mathbf{u}(t)|^p_{s_0,p}]. \end{aligned}$$

Similarly, by Lemma 3.1 there is $s_0 < s$ so that

$$\sum_k ||\eta_k(\sigma^i(t) - \sigma^i(t,x_k))\partial_i \mathbf{u}(t)||^p_{s,p}$$
$$= \sum_k ||\tilde{\eta}_k(\sigma^i(t) - \sigma^i(t,x_k))\eta_k \partial_i \mathbf{u}(t)||^p_{s,p}$$
$$\le C \sum_k [\sup_x |\tilde{\eta}_k(\sigma^i(t) - \sigma^i(t,x_k))|^p_Y |\eta_k \partial_i \mathbf{u}(t)|^p_{s,p} + |\eta_k \partial_i \mathbf{u}(t)|^p_{s_0,p}].$$

It follows by the assumptions, equation (4.12), Lemma 3.9, Corollary 3.10, and the interpolation theorem (see Lemma 6.7 in [2]) that for each ε there is $\kappa > 0$ and a constant $C = C(\varepsilon, \kappa, d, p, \delta, K)$ such that

$$\sum_k |\mathcal{S}(\mathbf{D}_k(\mathbf{u},t))|^p_{s-1,p} \le \varepsilon |\partial^2 \mathbf{u}(t)|^p_{s-1,p} + C(|\mathbf{u}(t,\cdot)|^p_{s-1,p} + |\mathbf{D}(\mathbf{0},t)|^p_{s-1,p}),$$

$$\sum_k ||\mathcal{S}(\mathbf{Q}_k(\mathbf{u},t,\cdot))||^p_{s,p} \le \varepsilon |\partial^2 \mathbf{u}(t)|^p_{s-1,p} + C(|\mathbf{u}(t)|^p_{s-1,p} + ||\mathbf{Q}(\mathbf{0},t)||^p_{s,p}).$$

Choosing ε sufficiently small and applying (4.12) and Theorem 4.3 to $\eta_k \mathbf{u}$ (it is a solution to (4.13)), we obtain that

(i) for each stopping time τ

$$\mathbf{E} \int_0^\tau |\partial^2 \mathbf{u}(t)|^p_{s-1,p}\, dt \le N \mathbf{E} \int_0^\tau (|\mathbf{u}(t)|^p_{s-1,p} + |\mathbf{D}(\mathbf{0},t)|^p_{s-1,p} + ||\mathbf{Q}(\mathbf{0},t)||^p_{s,p})\, dt,$$

where $N = N(p,d,\delta,K)$ does not depend on τ;

(ii) for each $T > 0$ and each stopping time $\tau \le T$

$$\mathbf{E} \sup_{t \le \tau} |\mathbf{u}(t)|^p_{s,p} \le N e^T \mathbf{E} \int_0^\tau (|\mathbf{u}(t)|^p_{s-1,p} + |\mathbf{D}(\mathbf{0},t)|^p_{s-1,p} + ||\mathbf{Q}(\mathbf{0},t)||^p_{s,p})\, dt. \tag{4.14}$$

Fix an arbitrary $\tau \le T$ such that

$$\mathbf{E}\left[\sup_{t \le \tau} |\mathbf{u}(t)|^p_{s,p} + \int_0^\tau (|\mathbf{u}(t)|^p_{s-1,p} + |\mathbf{D}(\mathbf{0},t)|^p_{s-1,p} + ||\mathbf{Q}(\mathbf{0},t)||^p_{s,p})\, dt\right] < \infty.$$

Then for each $t \le T$

$$\mathbf{E} \sup_{r \le t \wedge \tau} |\mathbf{u}(r)|^p_{s,p} \le N e^T \mathbf{E} \int_0^t \sup_{\bar{r} \le r \wedge \tau} |\mathbf{u}(\bar{r})|^p_{s,p}\, dr$$
$$+ \mathbf{E} \int_0^\tau |\mathbf{D}(\mathbf{0},t)|^p_{s-1,p} + ||\mathbf{Q}(\mathbf{0},t)||^p_{s,p}\, dt,$$

and the statement follows by Gronwall's inequality. □

Now we can prove the uniqueness of solutions of (4.1).

COROLLARY 4.5. *Let* A, A1(s,p)–A3(s,p) *hold,* $p \ge 2$. *Then there is at most one* $\mathbb{H}^s_p(\mathbf{R}^d)$*-solution to* (4.1) *in* $[[0,\tau]]$.

Proof. Assume that $\mathbf{u}_1, \mathbf{u}_2$ are $\mathbb{H}^s_p(\mathbf{R}^d)$-valued continuous solutions to (2.3) such that $\mathbf{P}$ a.s. for all t,

$$\int_0^{t \wedge \tau} |\partial^2 \mathbf{u}_l(r)|^p_{s-1,p}\, dr < \infty, \quad l = 1,2.$$

Then $\mathbf{v} = \mathbf{u}_2 - \mathbf{u}_1$ satisfies the equation

$$\begin{aligned}\partial_t \mathbf{v}(t,x) &= \partial_i(a^{ij}(t,x)\partial_j \mathbf{v}) + \mathbf{D}(\mathbf{v}+\mathbf{u}_1,t,x) - \mathbf{D}(\mathbf{u}_1,t,x)\\ &\quad + [\sigma^k(t,x)\partial_k \mathbf{v}(t,x) + \mathbf{Q}(\mathbf{v}+\mathbf{u}_1,t,x) - \mathbf{Q}(\mathbf{u}_1,t,x)]\cdot \dot{W},\\ \mathbf{v}(0,x) &= \mathbf{0}.\end{aligned}$$

Applying Lemma 4.4 to this equation and $\mathbf{v}$, we have $\mathbf{v} = \mathbf{0}$ by (4.10). □

To complete the proof of Theorem 2.3 we apply the standard method of continuity (see Theorem 5.1 in [2], Theorem 2 in [7]).

Proof of Theorem 2.3. The uniqueness follows by Corollary 4.5. So, we prove the existence of a solution to (4.1). Without any loss of generality we can assume $\mathbf{u}_0 = \mathbf{0}$ (see the proof of Theorem 5.1 in [2]), $\tau = \infty$. Then we introduce a parameter $\lambda \in [0,1]$ and consider the equation

$$\begin{aligned}\partial_t \mathbf{u}(t,x) &= \mathcal{S}\{\partial_i[\lambda\delta_{ij} + (1-\lambda)a^{ij}\partial_j \mathbf{u}] + \mathbf{D}(\mathbf{u},t,x)\}\\ &\quad + \mathcal{S}[(1-\lambda)\sigma^k \partial_k \mathbf{u} + \mathbf{Q}(\mathbf{u},t,x)]\cdot \dot{W},\\ \operatorname{div}\mathbf{u} &= 0,\end{aligned} \tag{4.15}$$

with zero initial condition. By Lemma 4.4 the a priori estimate (4.10) holds with the same constant C. Assume that for $\lambda = \lambda_0$ equation (4.15) for any $\mathbf{D}, \mathbf{Q}$ satisfying A3(s,p) has a unique continuous in t $\mathbb{H}^s_p$-valued solution such that $\mathbf{P}$ a.s. for all t,

$$\int_0^t |\partial^2 \mathbf{u}(r)|^p_{s-1,p}\, dr < \infty.$$

For other $\lambda \in [0,1]$ we rewrite (4.15) as

$$\begin{aligned}\partial_t \mathbf{u}(t,x) &= \mathcal{S}\{\partial_i[(\lambda_0\delta_{ij} + (1-\lambda_0)a^{ij})\partial_j \mathbf{u}] + \mathbf{D}(\mathbf{u},t,x)\\ &\quad + (\lambda-\lambda_0)\partial_i[(\delta_{ij}+a^{ij})\partial_j \mathbf{u}]\}\\ &\quad + \mathcal{S}[(1-\lambda_0)\sigma^i\partial_i \mathbf{u} + (\lambda-\lambda_0)\sigma^i\partial_i \mathbf{u} + \mathbf{Q}(\mathbf{u},t,x)]\cdot \dot{W},\\ \operatorname{div}\mathbf{u} &= 0\end{aligned}$$

and solve it by iterations. Define $\mathbf{u}_0 = \mathbf{0}$ and

$$\begin{aligned}\partial_t \mathbf{u}_{k+1}(t,x) &= \mathcal{S}\{\partial_i[(\lambda_0\delta_{ij} + (1-\lambda_0)a^{ij})\partial_j \mathbf{u}_{k+1}] + \mathbf{D}(\mathbf{u}_{k+1},t,x)\\ &\quad + (\lambda-\lambda_0)\partial_i[(\delta_{ij}+a^{ij})\partial_j \mathbf{u}_k]\}\\ &\quad + \mathcal{S}[(1-\lambda_0)\sigma^k\partial_k \mathbf{u}_{k+1} + (\lambda-\lambda_0)\sigma^i\partial_i \mathbf{u}_k + \mathbf{Q}(\mathbf{u}_{k+1},t,x)]\cdot \dot{W},\\ \operatorname{div}\mathbf{u} &= 0.\end{aligned} \tag{4.16}$$

So $\bar{\mathbf{u}}_{k+1} = \mathbf{u}_{k+1} - \mathbf{u}_k$ is a solution of the equation

$$\begin{aligned}\partial_t \bar{\mathbf{u}}_{k+1}(t,x) &= \mathcal{S}\{\partial_i[\lambda_0\delta_{ij} + (1-\lambda_0)a^{ij}\partial_j \bar{\mathbf{u}}_{k+1}] + \mathbf{D}(\mathbf{u}_k+\bar{\mathbf{u}}_{k+1},t,x) - \mathbf{D}(\mathbf{u}_k,t,x)\}\\ &\quad + \mathcal{S}\{(\lambda-\lambda_0)\partial_i[(\delta_{ij}+a^{ij})\partial_j \bar{\mathbf{u}}_k]\} + \mathcal{S}\{(\lambda-\lambda_0)\sigma^i\partial_i \bar{\mathbf{u}}_k\}\cdot \dot{W}\\ &\quad + \mathcal{S}[(1-\lambda_0)\sigma^k\partial_k \bar{\mathbf{u}}_{k+1} + \mathbf{Q}(\mathbf{u}_k+\bar{\mathbf{u}}_{k+1},t,x) - \mathbf{Q}(\mathbf{u}_k,t,x)]\cdot \dot{W},\end{aligned}$$

$\operatorname{div}\mathbf{u} = 0$. By our assumptions for each $T > 0$, there is a constant $C = C(d,p,\delta,K,T)$

such that for all stopping times $\tau \le T$

$$\begin{aligned}
&\mathbf{E}\left[\sup_{r\le\tau}|\bar{\mathbf{u}}_{k+1}(r)|^p_{s,p}+\int_0^\tau|\partial^2\bar{\mathbf{u}}_{k+1}(r)|^p_{s-1,p}\,dr\right]\\
&\le C'|\lambda-\lambda_0|^p\mathbf{E}\int_0^\tau(|\partial\bar{\mathbf{u}}_k(r)|_{s,p}+|\partial^2\bar{\mathbf{u}}_k(r)|^p_{s-1,p})\,dr\\
&\le C|\lambda-\lambda_0|^p\mathbf{E}\left[\sup_{r\le\tau}|\bar{\mathbf{u}}_k(r)|^p_{s,p}+\int_0^\tau|\partial^2\bar{\mathbf{u}}_k(r)|^p_{s-1,p}\,dr\right].
\end{aligned}$$

Fix an arbitrary stopping time $\tau \le T$ such that

$$I(\tau)=\mathbf{E}\left[\sup_{r\le\tau}|\mathbf{u}_1(r)|^p_{s,p}+\int_0^\tau|\partial^2\mathbf{u}_1(r)|^p_{s-1,p}\,dr\right]<\infty.$$

Notice that $\mathbf{u}_1$ and τ do not depend on λ (only on λ_0). Let $|\lambda-\lambda_0|<C^{-1/p}/2$. Then

$$\mathbf{E}\left[\sup_{r\le\tau}|\bar{\mathbf{u}}_{k+1}(r)|^p_{s,p}+\int_0^\tau|\partial^2\bar{\mathbf{u}}_{k+1}(r)|^p_{s-1,p}\,dr\right]^{1/p}\le(1/2)^kI(\tau)^{1/p},$$

and $(\mathbf{u}_k)$ is a Cauchy sequence on $[0,\tau]$. Therefore, there is a continuous in t $\mathbb{H}^s_p$-valued process $\mathbf{u}$ such that

$$\mathbf{E}\left[\sup_{r\le\tau}|\mathbf{u}_k(r)-\mathbf{u}(r)|^p_{s,p}+\int_0^\tau|\partial^2(\mathbf{u}_k(r)-\mathbf{u}(r))|^p_{s-1,p}\,dr\right]\to0$$

as $k\to\infty$. Obviously $\mathbf{u}$ is a solution to (4.15) on $[0,\tau]$. Since τ is any stopping time such that $I(\tau)$ is finite, it follows that we have a solution for any $|\lambda-\lambda_0|<C^{-1/p}/2$ (assuming we have one for λ_0). For $\lambda=1$ it exists by Theorem 4.3. So, in a finite number of steps starting with $\lambda=1$, we get to $\lambda=0$. This proves Theorem 2.3.

COROLLARY 4.6 (cf. Corollary 5.11 in [2]). *Let* A, A1(s,p)–A3(s,p), A1(s,q)–A3(s,q) *hold,* $p,q\ge2$, *and let* $|\mathbf{u}_0|_{s+1-2/p,p}+|\mathbf{u}_0|_{s+1-2/q,q}<\infty$ $\mathbf{P}$ *a.s. Then the* $\mathbb{H}^s_p$*-solution* $\mathbf{u}$ *from Theorem* 2.3 *is also an* $\mathbb{H}^s_q$*-solution, i.e., for each* $T>0$, *there is a constant* C *such that for each stopping time* $\tau\le T, A\in\mathcal{F}_0$,

$$\begin{aligned}
&\mathbf{E}1_A\left[\sup_{r\le\tau}|\mathbf{u}(r)|^l_{r,l}+\int_0^\tau|\partial^2\mathbf{u}(r)|^l_{s-1,l}\,dr\right]\\
&\quad\le C\mathbf{E}1_A\left[|\mathbf{u}_0|^l_{s+1,l}+\int_0^\tau(|\mathbf{D}(\mathbf{0},r)|^l_{s-1,l}+||\mathbf{Q}(\mathbf{0},r)||^l_{s,l})\,dr\right],
\end{aligned}$$

$l=p,q$.

Proof. We follow the lines of the proof of Theorem 2.3 by introducing the parameter $\lambda\in[0,1]$ and by considering (4.15). We can assume that $\mathbf{u}_0=\mathbf{0}$. The statement holds for $\lambda=1$ by Lemma 5.11 in [2] applied to each component of $\mathbf{u}$. If it is true for λ_0, then (4.16) defines a sequence $\mathbf{u}_k$ of $\mathbb{H}^s_p$-valued continuous processes that are $\mathbb{H}^s_q$-valued and continuous as well, and $\mathbf{P}$ a.s. for all t,

$$\int_0^t|\partial^2\mathbf{u}(r)|^l_{s-1,l}\,dr<\infty,\quad l=p,q.$$

For each $T>0$ there are constants $C_l=C(d,l,\delta,K,T)$, $l=p,q$, such that for all stopping times $\tau\le T$

$$\begin{aligned}&\mathbf{E}\left[\sup_{r\le\tau}|\bar{\mathbf{u}}_{k+1}(r)|_{s,l}^l+\int_0^\tau|\partial^2\bar{\mathbf{u}}_{k+1}(r)|_{s-1,l}^l\,dr\right]\\&\le C'|\lambda-\lambda_0|^p\mathbf{E}\int_0^\tau(|\partial\bar{\mathbf{u}}_k(r)|_{s,p}+|\partial^2\bar{\mathbf{u}}_k(r)|_{s-1,p}^p)\,dr\\&\le C_l|\lambda-\lambda_0|^p\mathbf{E}\left[\sup_{r\le\tau}|\bar{\mathbf{u}}_k(r)|_{s,l}^l+\int_0^\tau|\partial^2\bar{\mathbf{u}}_k(r)|_{s-1,l}^l\,dr\right],\end{aligned}$$

$l=p,q$. Fix an arbitrary stopping time $\tau\le T$ such that

$$I(\tau)=\mathbf{E}\left[\sup_{r\le\tau}(|\mathbf{u}_1(r)|_{s,p}^p+|\mathbf{u}_1(r)|_{s,q}^q)+\int_0^\tau(|\partial^2\mathbf{u}_1(r)|_{s-1,p}^p+|\partial^2\mathbf{u}_1(r)|_{s-1,q}^q)\,dr\right]<\infty.$$

Let $C=\max\{C_p,C_q\}$, $|\lambda-\lambda_0|<C^{-1/p}/2$. Then

$$\mathbf{E}\left[\sup_{r\le\tau}|\bar{\mathbf{u}}_{k+1}(r)|_{s,l}^l+\int_0^\tau|\partial^2\bar{\mathbf{u}}_{k+1}(r)|_{n-1,l}^l\,dr\right]^{1/p}\le(1/2)^kI(\tau)^{1/p},$$

$l=p,q$. Therefore, there is a continuous in t $\mathbb{H}_p^s\cap\mathbb{H}_q^s$-valued process $\mathbf{u}$ such that

$$\mathbf{E}\left[\sup_{r\le\tau}|\mathbf{u}_k(r)-\mathbf{u}(r)|_{s,l}^l+\int_0^\tau|\partial^2(\mathbf{u}_k(r)-\mathbf{u}(r))|_{s-1,l}^l\,dr\right]\to0,$$

$l=p,q$, and the statement follows. □

If s is large positive, assumption A3(s,p) is rarely satisfied even in the scalar case (see the example below). The following proposition helps to circumvent this problem in many important cases.

PROPOSITION 4.7. *Assume that for each* $\mathbf{v}\in\mathbb{H}_p^{s+1}$, $\mathbf{Q}(\mathbf{v},t)$ *is a predictable* $\mathbb{H}_p^{s+1}$*-valued process and* $\mathbf{D}(\mathbf{v},t)$ *is a predictable* $\mathbb{H}_p^s$*-valued process. Let* A, A1(s,p)–A3(s,p), A1$(s+1,p)$, A2$(s+1,p)$ *be satisfied,* $\mathbf{E}(|\mathbf{u}_0|_{s+2-2/p,p}^p)<\infty$, *and for all* $t>0,\mathbf{v}\in\mathbb{H}_p^{s+1}$,

$$\begin{aligned}||\mathbf{Q}(\mathbf{v},t)||_{s+1,p}&\le||\mathbf{Q}(\mathbf{0},t)||_{s+1,p}+C|\mathbf{v}|_{s+1,p},\\|\mathbf{D}(\mathbf{v},t)|_{s,p}&\le|\mathbf{D}(\mathbf{0},t)|_{s,p}+C|\mathbf{v}|_{s+1,p}.\end{aligned}$$

Suppose also that

$$\int_0^t(||\mathbf{Q}(\mathbf{0},r)||_{s+1,p}^p+|\mathbf{D}(\mathbf{0},r)|_{s,p}^p)\,dr<\infty$$

$\mathbf{P}$ *a.s. for all* t. *Then* (2.3) *has a unique continuous* $\mathbb{H}_p^{s+1}$*-solution.*

Moreover, for each $T>0$ *there is a constant* C *such that for each stopping time* $\tau\le T$,

$$\begin{aligned}&\mathbf{E}\left[\sup_{r\le\tau}|\mathbf{u}(r)|_{s+1,p}^p+\int_0^\tau|\partial^2\mathbf{u}(r)|_{s,p}^p\,dr\right]\\&\le C\mathbf{E}\left[|\mathbf{u}_0|_{s+2,p}^p+\int_0^\tau(|\mathbf{D}(\mathbf{0},r)|_{s,p}^p+||\mathbf{Q}(\mathbf{0},r)||_{s+1,p}^p)\,dr\right].\end{aligned}$$

Proof. Since the assumptions A, A1(s,p)–A3(s,p) are satisfied, the existence and uniqueness of $\mathbb{H}^s_p$-solution $\mathbf{u}$ is guaranteed by Theorem 2.3. By the same theorem, the linear equation

$$\begin{aligned}\partial_t \xi(t,x) &= \partial_i(a^{ij}(t,x)\partial_j\xi(t,x)) + \mathbf{D}(\mathbf{u},t,x)\\ &\quad + [\sigma^k(t,x)\partial_k\xi(t,x) + \mathbf{Q}(\mathbf{u},t,x)]\cdot \dot{W},\\ \xi(0,x) &= \mathbf{u}_0(x),\end{aligned}$$

has a unique $\mathbb{H}^{s+1}_p$-solution. Thus, $\xi = \mathbf{u}$ $\mathbf{P}$ a.s. Moreover, for each T there is a constant C such that for all stopping times $\tau \le T$,

$$\begin{aligned}&\mathbf{E}\left[\sup_{r\le t\wedge\tau}|\mathbf{u}(r)|^p_{s+1,p} + \int_0^{t\wedge\tau}|\partial^2\mathbf{u}(r)|^p_{s,p}\,dr\right]\\ &\le C\mathbf{E}\left[|\mathbf{u}_0|^p_{s+2,p} + \int_0^{t\wedge\tau}(|\mathbf{u}(r)|^p_{s+1,p} + |\mathbf{D}(\mathbf{0},r)|^p_{s,p} + ||\mathbf{Q}(\mathbf{0},r)||^p_{s+1,p})\,dr\right].\end{aligned}$$

Now the estimate of the statement follows by Gronwall's inequality. □

Example. Let us consider the following scalar equation:

$$\begin{aligned}\partial_t u &= \mathcal{S}[\Delta u + D(u)] + \mathcal{S}(u)\cdot\dot{W},\\ u(0,x) &= 0,\end{aligned}$$

where $W(t)$ is a one-dimensional Wiener process, $D(u) = \partial[f(u(x))] = \partial f(u(x))\partial u(x)$, and f is a scalar Lipschitz function on R^1. Then A3(1,p) would require the following estimate:

$$\begin{aligned}|D(u) - D(v)|_p &= |\nabla f(u(x))\partial u(x) - \nabla f(v(x))\partial v(x)|_p\\ &\le \varepsilon|u-v|_{2,p} + K_\varepsilon|u-v|_p,\end{aligned}$$

which is false in general even if ∇f is Lipschitz.

On the other hand, the assumptions of the proposition are satisfied for $s = 0$. Indeed,

$$|D(u)|_p = |\nabla f(u)\partial u|_p \le C|\partial u|_p,$$

where C is the Lipschitz constant of f.

Now, since ∂ is a bounded operator from $\mathbb{H}^s_p$ into $\mathbb{H}^{s+1}_p$, we have

$$\begin{aligned}&|D(u) - D(v)|_{-1,p} = |\partial[f(u)] - \partial[f(v)]|_{-1,p}\\ &\le C|f(u) - f(v)|_p \le C'|u-v|_p\\ &\qquad \le \varepsilon|u-v|_{1,p} + K_\varepsilon|u-v|_{-1,p}.\end{aligned}$$

(The latter inequality follows from Remark 5.5 in [2].) Thus assumption A3(0,p) is verified.

REFERENCES

[1] N. V. Krylov, *On L_p-theory of stochastic partial differential equations in the whole space,* SIAM J. Math. Anal., 27 (1996), pp. 313–340.

[2] N. V. Krylov, *An analytic approach to SPDEs,* in Stochastic Partial Differential Equations: Six Perspectives, Math. Surveys Monogr. 64, AMS, Providence, RI, 1999, pp. 185–242.

[3] R. Mikulevicius and B. L. Rozovskii, *On equations of stochastic fluid mechanics,* in Stochastics in Finite and Infinite Dimensions: In Honor of Gopinath Kallianpur, T. Hida, R. Karandikar, H. Kunita, B. Rajput, S. Watanabe, and J. Xiong, eds., Birkhauser Boston, Cambridge, MA, 2001, pp. 285–302.

[4] R. Mikulevicius and B. L. Rozovskii, *Stochastic Navier-Stokes equations. Propagation of chaos and statistical moments,* in J. L. Menaldi, E. Kofman, and A. Sulem, eds., Optimal Control and Partial Differential Equations: In Honour of Alain Bensoussan, IOS Press, Amsterdam, 2001, pp. 258–267.

[5] H. Triebel, *Theory of Function Spaces,* Birkhäuser-Verlag, Basel, 1983.

[6] E. Stein, *Singular Integrals and Differentiability Properties of Functions,* Princeton University Press, Princeton, NJ, 1970.

[7] R. Mikulevicius and B. Rozovskii, *A note on Krylov's L_p-theory for systems of SPDEs,* Electron. J. Probab., 6 (2001), paper 12, 35 pp.

[8] R. Mikulevicius and G. Valiukevicius, *On stochastic Euler equations in* $\mathbf{R}^d$, Electron. J. Probab., 5 (2000), paper 6, 20 pp.

SIAM J. MATH. ANAL.
Vol. 34, No. 1, pp. 142–172

WEAK STABILITY OF NONUNIFORMLY STABLE MULTIDIMENSIONAL SHOCKS*

JEAN-FRANÇOIS COULOMBEL†

Abstract. The aim of this paper is to investigate the linear stability of multidimensional shock waves that violate the uniform stability condition derived by Majda [*Mem. Amer. Math. Soc.*, 41 (1983)]. Two examples of such shock waves are studied: (1) planar Lax shocks in isentropic gas dynamics and (2) phase transitions in an isothermal van der Waals fluid. In both cases we prove an energy estimate on the resulting linearized system. Special attention is paid to the losses of derivatives arising from the failure of the uniform stability condition.

Key words. conservation laws, multidimensional shocks, phase transitions, stability of shock waves

AMS subject classifications. 35L45, 35L50, 35L65, 76L05, 76T10

PII. S0036141001392803

1. Introduction.

The stability of multidimensional shock waves in gas dynamics has been an active field of mathematical research since the late 1940's; see, e.g., [10, 12, 13, 19, 33]. The first results proved on this subject were giving some necessary conditions of stability by means of a normal modes analysis. In [21] (see also the review [22]), Lax formulated the definition of a shock wave for an arbitrary system of conservation laws in space dimension one: the definition was also dictated by some kind of "stability" argument. More precisely, the number of characteristics impinging on the shock front curve is imposed by the size of the system in order to avoid under- (or over-) determinacy of the resulting free boundary problem. Regarding ideal gas dynamics, this definition is known to be equivalent to the requirement that the physical entropy increases upon crossing the shock front curve; see [10].

Using the extensive study of initial boundary value problems for linear hyperbolic systems (see, e.g., [16, 17, 20]), Majda succeeded in the early 1980's in deriving a necessary and sufficient strong stability condition for multidimensional shock waves [25]. The resulting estimates on the linearized problem enabled him to prove a nonlinear existence theorem [24]. We also refer to [26, 37] for a general overview of the method and its application to isentropic gas dynamics. It is worth noting that a different approach developed at the same time by Blokhin [6, 7] gave rise to similar results. However, Majda's approach, which has been slightly improved in [27, 30] by using the new ideas of paradifferential calculus introduced by Bony and Meyer, seems appropriate to our purpose, and we shall adopt it for our analysis.

In the study of initial boundary value problems for linear hyperbolic systems, many physically relevant boundary data are found to violate the uniform stability condition, namely the so-called Kreiss–Lopatinskii condition. A list of such boundary conditions for physical systems can be found in [11]. Nevertheless, many authors have overcome this difficulty in various cases by using particular properties of the involved (linear or nonlinear) system; see, e.g., [3, 15, 35] for results on fluid dynamics and [29, 34] for results on elastodynamics. In a more general setting, Ohkubo and

*Received by the editors July 24, 2001; accepted for publication (in revised form) March 22, 2002; published electronically September 5, 2002.
http://www.siam.org/journals/sima/34-1/39280.html

†UMPA, CNRS-UMR 5669, École Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France (jfcoulom@umpa.ens-lyon.fr).

Shirota derived in [31] a sufficient condition for linear initial boundary value problems to ensure L^2 well-posedness with respect to the "interior source term." (Initial and boundary data are homogeneous.)

Although Majda's result has the great advantage of dealing with any system of conservation laws, examples of multidimensional shocks are not that numerous, and the verification of the uniform stability condition often gives rise to very tedious computations. However, such verification can be carried out for the system of gas dynamics. Two cases of nonuniformly stable shocks arise and motivate the present study. The first example, which is briefly addressed in [25], is the one of planar Lax shocks in isentropic gas dynamics that violate Majda's inequality (see [25, p. 10]). This inequality is recalled in section 2. The second example comes from the theory of phase transitions in isothermal van der Waals fluids. These planar discontinuities are undercompressive shocks. They require an additional jump relation to select the relevant ones. Various admissibility criteria have been proposed over the last two decades; see [39] for phase transitions in the context of gas dynamics or [38, 40] and references therein for phase transitions in the context of elastodynamics. We base our analysis on the viscosity-capillarity criterion proposed in [39] under the assumption that the viscosity coefficient is neglected and taken to be zero. In other words, the additional jump relation is written as a generalized equal area rule. It has been shown in [4] that the uniform stability condition is violated because of surface waves. (Taking viscosity into account would yield uniform stability; see [5].) It is worth noting that the failure of the uniform stability condition in isentropic gas dynamics can rise only from the appearance of boundary waves (but we shall get back to this in the next sections); for a precise statement of the distinction between these two types of waves, we refer the reader to the very nice survey [11].

The purpose of the paper is the derivation of a complete energy estimate on the linearized system resulting from the study of these two problems. Since the *classical* energy estimate is known to be equivalent to the uniform stability condition, as proved in [25], losses of derivatives are to be expected. As shown in Theorems 3.5 and 4.5, and this is no real surprise, losses of derivatives are more severe when boundary waves occur than when surface waves occur. We point out that this kind of phenomenon had already been mentioned in previous works [11, 34]. Despite the impossibility of using some "dissipativeness" arguments on the boundary conditions in our context, we shall see that the derivation of an energy estimate can be carried out by a suitable modification in the ordinary construction of a Kreiss symmetrizer. This point will be emphasized in both problems we shall detail.

This paper is divided as follows. In section 2, we recall Majda's method for multidimensional shock waves and introduce some notations. Note that Lax shocks for isentropic Euler equations are uniformly stable in one space dimension, and we shall therefore deal with two- or three-dimensional problems. (The one-dimensional case is treated in [23].) We warn the reader that many calculations cannot be reproduced here to avoid overloading the paper, and we shall often refer to previous works on this subject where some details are available. However, special attention will be paid to detailing the normal modes analysis on which relies the entire construction of the symbolic symmetrizer. In section 3, we treat the first example, i.e., nonuniformly stable Lax shocks for isentropic Euler equations. We show in section 4 how the method developed in section 3 applies in the study of phase transitions in a van der Waals fluid and even gives slightly better results. Once again, we shall focus on two- or three-dimensional problems, since phase transitions are known to be uniformly

stable in one space dimension, and their existence has already been studied in [14]. Section 5 is devoted to the proof of several technical lemmas used in the construction of Kreiss symmetrizers. Eventually, we make in section 6 some general remarks on the possible advances for these two problems.

2. General considerations. We study the Euler equations governing the motion of an inviscid isentropic fluid in $\mathbb{R}^d$:

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho \mathbf{u}) = 0 \,, \\ \partial_t (\rho \mathbf{u}) + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \nabla p = 0 \,. \end{cases} \tag{2.1}$$

We have adopted the following standard notations that will be used throughout this paper: ρ denotes the density, $\mathbf{u}$ the velocity field, and c the sound speed given by the pressure law $p(\rho)$ that the fluid is assumed to obey,

$$c(\rho) = \sqrt{p'(\rho)} \,.$$

Since smooth solutions generally develop singularities in finite time, we look for particular weak solutions of the form of functions which are smooth on both sides of a (variable) hypersurface of $\mathbb{R}^d$. A first step in the proof of the existence of such solutions is the study of the linear stability of piecewise constant solutions defined by a relation of the form

$$\overline{U} = \begin{cases} U_l = (\rho_l, \mathbf{u}_l) & \text{if } x \cdot \nu < \sigma t, \\ U_r = (\rho_r, \mathbf{u}_r) & \text{if } x \cdot \nu > \sigma t. \end{cases}$$

Such a function $\overline{U}$ is a weak solution of the Euler equations (2.1) if and only if it satisfies the Rankine–Hugoniot jump relations which can be written in the following way:

$$\begin{cases} \rho_r(\mathbf{u}_r \cdot \nu - \sigma) = \rho_l(\mathbf{u}_l \cdot \nu - \sigma) =: j \,, \\ j[\mathbf{u}] + [p]\nu = 0 \,. \end{cases} \tag{2.2}$$

We consider dynamical discontinuities and thus assume that the mass transfer j across the hyperplane $\{x \cdot \nu = \sigma t\}$ is not zero. By symmetry arguments, one can therefore assume $j > 0$. We first assume that $\overline{U}$ defines a compressive 1-Lax shock or, in other words, that the following inequalities hold:

$$M_r = \frac{\mathbf{u}_r \cdot \nu - \sigma}{c(\rho_r)} < 1, \quad M_l = \frac{\mathbf{u}_l \cdot \nu - \sigma}{c(\rho_l)} > 1, \quad \text{and} \quad \rho_r > \rho_l \,.$$

Note that the above assumptions immediately imply that the shock is noncharacteristic: the propagation speed of the interface σ is different from the characteristic speeds of system (2.1) on both sides of the interface. With the above notations, we have the following statement.

PROPOSITION 2.1 (Majda [25]). *The shock $\overline{U}$ is uniformly stable if and only if*

$$M_r^2 \left(\frac{\rho_r}{\rho_l} - 1 \right) < 1 \,. \tag{2.3}$$

If inequality (2.3) *does not hold, then the shock $\overline{U}$ is only weakly stable.*

Inequality (2.3) holds as long as p is a convex function of the density ρ which is the case for the classical gamma-law but not for more complicated laws (like, for instance, an isothermal van der Waals pressure law). We shall investigate in section 3 the case where the opposite *strict* inequality holds. We shall also detail why the equality case cannot be treated by the techniques used in this paper.

If we now assume that p is a nonmonotone function of ρ (this hypothesis can be viewed as a model of isothermal liquid-vapor phase transitions; see [18]), it is known that subsonic discontinuities can appear for which we have

$$M_r = \frac{\mathbf{u}_r \cdot \nu - \sigma}{c(\rho_r)} < 1, \quad M_l = \frac{\mathbf{u}_l \cdot \nu - \sigma}{c(\rho_l)} < 1, \quad \text{and} \quad \rho_r > \rho_l \,.$$

Such inequalities occur if p is, for instance, given by an isothermal van der Waals pressure law with a temperature below the so-called critical temperature (see [4, 5, 39]). To avoid the natural instability of $\overline{U}$ with respect to small perturbations, one needs to specify an additional jump relation to the Rankine–Hugoniot conditions. The analysis developed in section 4 is based on the capillarity criterion proposed in [39]. (The admissibility criterion proposed in [40] is the analogue for elastodynamics, and the main idea governing both criteria is that there is no entropy dissipation upon crossing the shock.) Together with the Rankine–Hugoniot conditions, this criterion requires that $\overline{U}$ satisfies the generalized equal area rule

$$\int_{v_r}^{v_l} p(v)\, dv \;=\; (v_l - v_r)\, \frac{p(v_r) + p(v_l)}{2}\,, \tag{2.4}$$

where $v = 1/\rho$ is the specific volume of the fluid. Such phase transitions which differ from Maxwell equilibrium states are noncharacteristic.

We are now able to develop Majda's method to study the linear stability of such multidimensional shocks. First note that, by a change of observer, one can always assume that the unit vector ν is the last vector of the canonical basis of $\mathbb{R}^d$. Since the mass transfer j is not zero, equations (2.2) show that the tangential components of the velocity are the same on both sides of the shock front curve. Performing another change of observer one can assume from now on that

$$(\mathbf{u}_1^r, \dots, \mathbf{u}_{d-1}^r) = (\mathbf{u}_1^l, \dots, \mathbf{u}_{d-1}^l) = \mathbf{0} \quad \text{and} \quad \sigma = 0\,,$$

which is of no consequence on the stability of the particular solution $\overline{U}$. Note that these operations yield a simplified expression of the mass transfer j across the interface (defined by system (2.2)): $j = \rho_r u_r = \rho_l u_l$.

We adopt in all that follows the following notations: all space vectors x in $\mathbb{R}^d$ are decomposed as $x = (y, x_d)$, where y is a vector in $\mathbb{R}^{d-1}$ and x_d is a scalar. Similarly, all velocity vectors $\mathbf{u}$ are decomposed as $\mathbf{u} = (\check{u}, u)$, where $\check{u} \in \mathbb{R}^{d-1}$ is the tangential part of the velocity and $u \in \mathbb{R}$ is the normal velocity.

We are now led to search a weak solution U of (2.1) defining a compressive 1-Lax shock (or an admissible phase transition) across a smooth hypersurface $\Sigma(t) = \{x_d = \varphi(t, y)\}$ close to the hyperplane $\{x_d = 0\}$. Since $\Sigma(t)$ is part of the unknowns of the problem, one first fixes the front by the following well-known transformation in free boundary problems:

$$\big(U :\, (t, y, x_d) \longrightarrow \mathbb{R}^N\,\big) \longrightarrow (U_\pm :\, (t, y, z) \longmapsto U(t, y, \varphi(t, y) \pm z)\,)\,,$$

both applications $U_+ = (\rho_+, \mathbf{u}_+)$ and $U_- = (\rho_-, \mathbf{u}_-)$ being defined on the same half-space $\{z > 0\}$. The quasi-linear form of Euler equations is linearized on both sides

of $\Sigma(t)$ around the piecewise constant solution $\overline{U}$ (see [25, 37]). The resulting linear system reads

$$\begin{cases} \partial_t U_+ + \sum_{j=1}^{d-1} A_j(U_r)\,\partial_{x_j} U_+ + A_d(U_r)\,\partial_z U_+ = f_+\,, \\ \partial_t U_- + \sum_{j=1}^{d-1} A_j(U_l)\,\partial_{x_j} U_- - A_d(U_l)\,\partial_z U_- = f_-\,, \end{cases} \tag{2.5}$$

where $A_j(U_{r,l})$ are $(d+1)\times(d+1)$ matrices corresponding to the quasi-linear form of isentropic Euler equations; see [9, 10, 36].

The linearization of the jump conditions across the interface $\Sigma(t)$ yields the boundary conditions on $\{z=0\}$. When one deals with a compressive Lax shock, the jump conditions are nothing but the Rankine–Hugoniot relations, and their linearized form reads

$$\begin{aligned} u_r\rho_+ + \rho_r u_+ - u_l\rho_- - \rho_l u_- - [\rho]\partial_t\varphi = g_1\,, \\ \rho_r u_r \check{u}_+ - \rho_l u_l \check{u}_- - [p]\nabla_y\varphi = \check{g}\,, \\ (u_r^2+c_r^2)\rho_+ + 2\rho_r u_r u_+ - (u_l^2+c_l^2)\rho_- - 2\rho_l u_l u_- = g_{d+1}\,. \end{aligned} \tag{2.6}$$

When one deals with a subsonic phase transition in a van der Waals fluid, the complete boundary conditions for the linearized problem are obtained by linearizing (2.4) and adding this new relation to the linearized Rankine–Hugoniot relations (2.6). The complete set of boundary conditions in this case reads

$$\begin{aligned} u_r\rho_+ + \rho_r u_+ - u_l\rho_- - \rho_l u_- - [\rho]\partial_t\varphi = g_1\,, \\ \rho_r u_r \check{u}_+ - \rho_l u_l \check{u}_- - [p]\nabla_y\varphi = \check{g}\,, \\ (u_r^2+c_r^2)\rho_+ + 2\rho_r u_r u_+ - (u_l^2+c_l^2)\rho_- - 2\rho_l u_l u_- = g_{d+1}\,, \\ c_r^2\frac{\rho_+}{\rho_r} + u_r u_+ - c_l^2\frac{\rho_-}{\rho_l} - u_l u_- - [u]\partial_t\varphi = g_{d+2}\,. \end{aligned} \tag{2.7}$$

It is now clear that, even though both examples rise from two different research areas, they are exactly of the same kind. In both cases, we are led to study a non-standard mixed initial boundary value problem

$$\begin{cases} \partial_t U + \sum_{j=1}^{d-1} \mathcal{A}_j\,\partial_{x_j} U + \mathcal{A}_d\,\partial_z U = f & \text{for } z>0, \\ \partial_t\varphi\, b_0 + \sum_{j=1}^{d-1} \partial_{x_j}\varphi\, b_j + M\,U = g & \text{for } z=0. \end{cases} \tag{2.8}$$

The boundary conditions for the study of compressive Lax shocks are given by (2.6), and the boundary conditions for the study of subsonic phase transitions are given by (2.7). To write system (2.8), we have let

$$U = \begin{pmatrix} U_+ \\ U_- \end{pmatrix}, \quad f = \begin{pmatrix} f_+ \\ f_- \end{pmatrix}, \quad g = \begin{pmatrix} g_1 \\ \check{g} \\ g_{d+1} \end{pmatrix}, \quad \text{or } g = \begin{pmatrix} g_1 \\ \check{g} \\ g_{d+1} \\ g_{d+2} \end{pmatrix},$$

$$\mathcal{A}_j = \begin{pmatrix} A_j(U_r) & \mathbf{0} \\ \mathbf{0} & A_j(U_l) \end{pmatrix} \text{ for } 1\le j\le d-1, \quad \mathcal{A}_d = \begin{pmatrix} A_d(U_r) & \mathbf{0} \\ \mathbf{0} & -A_d(U_l) \end{pmatrix}.$$

In both examples, M represents the matrix of the linearized jump conditions (Rankine–Hugoniot relations and the generalized equal area rule in the case of phase transitions). The vectors $b_0, \dots, b_{d-1}$ come from (2.6) and (2.7). They belong to $\mathbb{R}^{d+1}$ in the study of Lax shocks, while they belong to $\mathbb{R}^{d+2}$ in the study of subsonic phase transitions.

The derivation of an energy estimate for system (2.8) relies on the introduction of a positive weight γ (see [20, 25]). More precisely, we perform a change of unknown functions

$$v(t,y,z) = e^{-\gamma t}\, U(t,y,z) \quad \text{and} \quad \psi(t,y) = e^{-\gamma t}\, \varphi(t,y)\,,$$

where γ is a nonnegative parameter. We now perform a Fourier transform in the variables t and y. (The corresponding dual variables will be respectively denoted δ and η.) These operations yield the system of ordinary differential equations

$$\begin{cases} \dfrac{dV}{dz} = \mathcal{A}(\delta,\eta,\gamma)\, V(z) + F & \text{for } z > 0, \\ \chi\, b(\delta,\eta,\gamma) + M\, V(0) = G & \text{for } z = 0, \end{cases} \tag{2.9}$$

with

$$\mathcal{A}(\delta,\eta,\gamma) = -\mathcal{A}_d^{-1} \left(\tau + i \sum_{j=1}^{d-1} \eta_j \mathcal{A}_j \right) \quad \text{and} \quad b(\delta,\eta,\gamma) = \tau b_0 + i \sum_{j=1}^{d-1} \eta_j b_j\,.$$

For convenience we have let $\tau = \gamma + i\delta$. Note that inverting $\mathcal{A}_d$ is legitimate, since the shock is in both examples noncharacteristic. We now turn to the description of the method: in both examples, we show that the boundary conditions in problem (2.9) can be rewritten so that χ appears only in the last scalar boundary condition. The remaining part of the work consists of deriving an a priori estimate on the resulting initial boundary value problem for U where the boundary conditions take the form of a pseudodifferential operator.

Because of the decoupled nature of system (2.5), it is clear that matrix $\mathcal{A}(\delta,\eta,\gamma)$ has a block diagonal structure: its first block corresponds to the linearized system ahead of the shock, and its second block corresponds to the linearized system before the shock (see [25, 26, 37]). The eigenmodes of the first block are $\omega_2^r = -\tau/u_r$, and the roots of the second order polynomial equation are

$$(\tau + u_r\, \omega)^2 = c_r^2 (\omega^2 - |\eta|^2)\,. \tag{2.10}$$

In a similar way, the eigenmodes of the second block are $\omega_2^l = \tau/u_l$, and the roots of the second order polynomial equation are

$$(\tau - u_l\, \omega)^2 = c_l^2 (\omega^2 - |\eta|^2)\,. \tag{2.11}$$

We briefly analyze the eigenmodes of $\mathcal{A}$ and begin with the eigenmodes of the first block. In both problems analyzed in sections 3 and 4, the shock $\overline{U}$ is subsonic with respect to the right state ($M_r < 1$). It is clear that ω_2^r is of negative real part when τ has positive real part (that is, when γ is positive). Moreover, (2.10) has one root ω_3^r of negative real part when τ has positive real. The other root of (2.10) is denoted ω_1^r and has positive real part when τ has positive real part. The parametrization of the corresponding eigenspaces, which we use in sections 3 and 4, can be found in [4, 37]. One crucial property of the eigenmodes $\omega_{1,3}^r$ is that they can be extended up

to imaginary values of τ. Note that ω_3^r has negative real part if $|\tau| < |\eta|\sqrt{c_r^2 - u_r^2}$ and is purely imaginary if $|\tau| \geq |\eta|\sqrt{c_r^2 - u_r^2}$.

In the case of a compressive 1-Lax shock, that is, when the shock is supersonic with respect to the left state, then the second dynamical system does not give any contribution to the stable subspace $\mathcal{E}^-$ of $\mathcal{A}$. Indeed, ω_2^l is of positive real part when τ has positive real part. Furthermore, (2.11) has two roots ω_1^l and ω_3^l of positive real part when τ has positive real part. One easily checks that the continuous extension of ω_1^l and ω_3^l for purely imaginary values of τ are always distinct.

In the case of a subsonic phase transition, (2.11) has the same behavior as (2.10). More precisely, (2.11) has exactly one root ω_1^l of negative real part when τ has positive real part. The other root of (2.11) is denoted ω_3^l. It has positive real part when τ has positive real part. When τ is a purely imaginary number, ω_1^l has negative real part if $|\tau| < |\eta|\sqrt{c_l^2 - u_l^2}$ and is purely imaginary if $|\tau| \geq |\eta|\sqrt{c_l^2 - u_l^2}$.

3. Nonuniformly stable shocks in gas dynamics. We begin by describing the failure of the uniform stability condition for compressive Lax shocks in isentropic gas dynamics. Let $\overline{U}$ define a compressive 1-Lax shock for isentropic Euler equations (2.1) as described in the previous section. We study the nonstandard initial boundary value problem (2.8) with boundary conditions given by (2.6). We assume that $\overline{U}$ violates Majda's inequality (2.3) in the following way:

$$M_r^2 \left(\frac{\rho_r}{\rho_l} - 1 \right) > 1 \,.$$

Note that the previous simplifications imply that this inequality is equivalent to

$$u_r u_l > c_r^2 + u_r^2 \,. \tag{3.1}$$

This remark will be useful to complete the proof of Lemma 3.3. Under the assumptions made on $\overline{U}$, the normal modes analysis of problem (2.9) is summarized in the following result.

LEMMA 3.1. *There exists a positive number V_1 such that, for all $(\delta, \eta, \gamma) \in \mathbb{R}^{d+1}$ satisfying $\gamma \geq 0$ and $(\delta, \gamma) \neq (\pm i V_1 |\eta|, 0)$, one has*

$$\big\{ (Z, \chi) \in \mathcal{E}^-(\delta, \eta, \gamma) \times \mathbb{C} \ \text{ so that (s.t.) } \ \chi\, b(\delta, \eta, \gamma) + M\, Z = 0 \big\} = \{0\} \,,$$

and, for $\eta \neq 0$, the set

$$\big\{ (Z, \chi) \in \mathcal{E}^-(\pm V_1 |\eta|, \eta, 0) \times \mathbb{C} \ \text{ s.t. } \ \chi\, b(\pm V_1 |\eta|, \eta, 0) + M\, Z = 0 \big\}$$

is a one-dimensional subspace of $\mathbb{C}^{2d+3}$.

By definition, V_1^2 is the smallest root of the polynomial

$$P_1(X) = (c_r^2 - u_r^2)(X^2 + u_r^2 u_l^2) + \big[4 u_r^2 c_r^2 - 2 u_r u_l (c_r^2 + u_r^2) \big]\, X \,,$$

which has two real positive roots under assumption (3.1). *(The greatest is denoted V_2^2.) Furthermore, we have*

$$c_r^2 - u_r^2 < V_1^2 < u_r u_l \frac{c_r^2 - u_r^2}{c_r^2 + u_r^2} < V_2^2 \,.$$

Proof. This is a basic extension of the calculations already done in [25] (which can also be found in [37]). First of all, we note that the stable subspace of the dynamical system

$$\frac{dV}{dz} = \mathcal{A}(\delta, \eta, \gamma)\, V$$

consists of all vectors $Z = (Z_r, Z_l)$ such that

$$(u_r\tau - (c_r^2 - u_r^2)\omega_3^r, \rho_r u_r i\eta^T, -\rho_r\tau) \cdot Z_r = 0 \quad \text{and} \quad Z_l = 0\,.$$

With this parametrization of the stable subspace, one easily computes the Lopatinskii determinant associated with (2.9):

$$\Delta(\delta, \eta, \gamma) = \rho_r^d u_r^{d-1} \left[(c_r^2 - u_r^2)[p]|\eta|^2 + (c_r^2 + u_r^2)[\rho]\tau^2 + 2u_r[\rho]\tau a_3^r\right]\,,$$

where we have let $a_3^r = u_r\tau - (c_r^2 - u_r^2)\omega_3^r$. It is clear that $\Delta(\delta, 0, \gamma)$ does not vanish for any $(\delta, \gamma) \neq (0, 0)$. One can therefore factor the expression of $\Delta(\delta, \eta, \gamma)$ by $|\eta|^2$ and use the reduced variables

$$V = \frac{\tau}{i|\eta|}, \quad A_3^r = \frac{a_3^r}{i|\eta|}\,.$$

Some simplifications using the Rankine–Hugoniot relations lead to the expression

$$\Delta(\delta, \eta, \gamma) = \rho_r^d u_r^{d-1}|\eta|^2[\rho]\left[(c_r^2 - u_r^2)u_r u_l - (c_r^2 + u_r^2)V^2 - 2u_r V A_3^r\right]\,.$$

Let $\mathcal{R}$ denote the complex square root mapping defined by

$$\begin{aligned}\mathcal{R}\ : \mathbb{C} \setminus \mathbb{R}_+ &\longrightarrow \{\zeta \in \mathbb{C} \text{ s.t. } \operatorname{Im}\zeta > 0\},\\ w &\longmapsto \mathcal{R}(w) \quad \text{with} \quad \mathcal{R}(w)^2 = w\,.\end{aligned}$$

Then analyzing equation (2.10) shows that for $\gamma > 0$ (or, equivalently, for V of negative imaginary part) we have

$$A_3^r = -c_r\mathcal{R}(V^2 - (c_r^2 - u_r^2)),$$

and therefore, if the Lopatinskii determinant vanishes at some point (τ, η), V^2 has to be a root of the polynomial P_1 defined in the lemma. Note that the assumption (3.1) made on the shock $\overline{U}$ implies that P_1 has two distinct positive roots V_1^2 and V_2^2 that satisfy the properties given in the lemma. This already proves that the possible zeros of $\Delta(\delta, \eta, \gamma)$ have to satisfy

$$\eta \neq 0\,,\ \gamma = 0 \quad \text{and} \quad \delta^2 > (c_r^2 - u_r^2)|\eta|^2\,,$$

and those requirements imply that V is a real number such that $V^2 > c_r^2 - u_r^2$. One therefore has to extend the previous definition of A_3^r to such values of V. This is achieved by using the Cauchy–Riemann relations on holomorphic functions (see [4, 37] for the details):

$$\begin{cases} A_3^r = c_r\sqrt{V^2 - (c_r^2 - u_r^2)} & \text{if } V > \sqrt{c_r^2 - u_r^2},\\ A_3^r = -c_r\sqrt{V^2 - (c_r^2 - u_r^2)} & \text{if } V < -\sqrt{c_r^2 - u_r^2}. \end{cases}$$

Furthermore, the previous analysis shows that $\Delta(\delta,\eta,0)$ vanishes if and only if

$$\begin{cases} 2u_rc_rV\sqrt{V^2-(c_r^2-u_r^2)} = -(c_r^2+u_r^2)V^2 + u_ru_l(c_r^2-u_r^2) & \text{if } V > \sqrt{c_r^2-u_r^2}, \\ 2u_rc_rV\sqrt{V^2-(c_r^2-u_r^2)} = (c_r^2+u_r^2)V^2 - u_ru_l(c_r^2-u_r^2) & \text{if } V < -\sqrt{c_r^2-u_r^2}, \end{cases}$$

and these relations imply $P_1(V^2)=0$.

If the Lopatinskii determinant vanishes at $V=V_2$, then we must have

$$2u_rc_rV_2\sqrt{V_2^2-(c_r^2-u_r^2)} = -(c_r^2+u_r^2)V_2^2 + u_ru_l(c_r^2-u_r^2)\,.$$

However, the left-hand term of the equality is positive, and the right-hand term is negative. Therefore the Lopatinskii determinant cannot vanish at $V=V_2$ (and neither at $V=-V_2$ by a similar argument). Since $P_1(V_1^2)=0$ we have

$$2u_rc_rV_1\sqrt{V_1^2-(c_r^2-u_r^2)} = -(c_r^2+u_r^2)V_1^2 + u_ru_l(c_r^2-u_r^2)\,,$$

because both terms in the equality are positive. Therefore the Lopatinskii determinant vanishes at $V=V_1$ (and similarly at $V=-V_1$). This completes the proof of the existence and the characterization of points where the uniform stability condition fails. The last assertion on the dimension of the corresponding kernel follows directly from the shape of the boundary conditions (2.6). □

Note that if in the special case $u_ru_l = c_r^2+u_r^2$, then $P_1(c_r^2-u_r^2)=0$. In other words, the uniform stability condition fails exactly at the points where (2.10) has a double root. At such points, the symbol $\mathcal{A}$ is not diagonalizable, and a 2×2 Jordan block arises in the reduction of $\mathcal{A}$ which is used to construct a Kreiss symmetrizer (see the proof of Proposition 3.4). At the present time, we have not been able to overcome this difficulty. This case is left to a future work.

3.1. Elimination of the front. The first step in the derivation of an energy estimate for the mixed problem (2.8) is to work in the Fourier space and to isolate the front χ in the last boundary condition for problem (2.9). This operation can be summarized in the following terms.

LEMMA 3.2. *There exists a C^∞ mapping Q defined on the half-space $\mathbb{R}^d\times\mathbb{R}^+\setminus\{0\}$, homogeneous of degree 0, with values in the set of square $(d+1)\times(d+1)$ invertible matrices such that, for all $X\in\mathbb{R}^d\times\mathbb{R}^+\setminus\{0\}$, the first d components of the vector $Q(X)\,b(X)$ vanish.*

Proof. The Rankine–Hugoniot jump relations together with (2.6) yield the relations

$$b(\delta,\eta,\gamma) = \begin{pmatrix} -\tau[\rho] \\ -iu_ru_l[\rho]\eta \\ 0 \end{pmatrix} \text{ if } d=2 \quad \text{and} \quad b(\delta,\eta,\gamma) = \begin{pmatrix} -[\rho]\tau \\ -iu_ru_l[\rho]\eta_1 \\ -iu_ru_l[\rho]\eta_2 \\ 0 \end{pmatrix} \text{ if } d=3.$$

To preserve the homogeneity of the physical quantities we handle in the calculations, we fix a reference velocity $\widetilde{V}$ and a reference frequency $\widetilde{\gamma}$, and we define Σ_+ as the hemisphere

$$\Sigma_+ = \left\{ (\delta,\eta,\gamma)\in\mathbb{R}^d\times\mathbb{R}_+ \text{ s.t. } \gamma^2+\delta^2+\widetilde{V}^2|\eta|^2 = \widetilde{\gamma}^2 \right\}.$$

We first define the mapping Q on the hemisphere Σ_+ and then extend it as a homogeneous mapping of degree 0. One easily checks that, for $d = 2$, the matrix

$$Q(\delta,\eta,\gamma) = \begin{pmatrix} 0 & 0 & 1 \\ iu_r u_l \eta & -\tau & 0 \\ u_r u_l \overline{\tau} & -i\widetilde{V}^2\eta & 0 \end{pmatrix}$$

satisfies all required properties. For $d = 3$, one can choose, for instance,

$$Q(\delta,\eta,\gamma) = \begin{pmatrix} 0 & 0 & 0 & 1 \\ iu_r u_l \eta_1 & -\tau & 0 & 0 \\ iu_r u_l \eta_2 & 0 & -\tau & 0 \\ u_r u_l \overline{\tau} & -i\widetilde{V}^2\eta_1 & -i\widetilde{V}^2\eta_2 & 0 \end{pmatrix}$$

which also satisfies all required properties. This completes the proof. □

We can therefore write boundary conditions for the linearized problem (2.8) in the equivalent way

$$\begin{pmatrix} B(\delta,\eta,\gamma) \\ \ell(\delta,\eta,\gamma) \end{pmatrix} V(0) + \chi \begin{pmatrix} \mathbf{0}_d \\ \alpha(\delta,\eta,\gamma) \end{pmatrix} = Q(\delta,\eta,\gamma)\, G\,,$$

where $\alpha(\delta,\eta,\gamma)$ is given by

$$\alpha(\delta,\eta,\gamma) = -u_r u_l [\rho] \widetilde{\gamma} \sqrt{\gamma^2 + \delta^2 + \widetilde{V}^2|\eta|^2} \neq 0\,,$$

and this relation holds for $d = 2$ and $d = 3$.

Lemma 3.1 ensures that the restriction of $B(X)$ to the stable subspace $\mathcal{E}^-(X)$ is invertible except at the points X where the uniform stability condition fails. We thus have to study the behavior of the restriction of $B(X)$ to the stable subspace $\mathcal{E}^-$ in the neighborhood of those points. Lemma 3.3 asserts that the Lopatinskii determinant vanishes *at order* 1 or, in other words, that the roots exhibited in Lemma 3.1 are simple.

For all vectors Z belonging to the stable subspace $\mathcal{E}^-$ we denote by Z_3^r and Z_2^r the components of Z on the eigenspaces associated with the eigenmodes ω_3^r and ω_2^r. In other words, we decompose Z as

$$Z = \begin{pmatrix} Z_r \\ \mathbf{0}_{d+1} \end{pmatrix} \quad \text{with} \quad Z_r = Z_3^r \begin{pmatrix} \rho_r(\tau + u_r\,\omega_3^r) \\ -c_r^2\, i\eta \\ -c_r^2\,\omega_3^r \end{pmatrix} + \begin{pmatrix} 0 \\ -\omega_2^r Z_2^r \\ i\eta \cdot Z_2^r \end{pmatrix}.$$

Then we have the following microlocal estimate.

LEMMA 3.3. *There exists a neighborhood $\mathcal{V}$ of $(V_1|\eta|,\eta,0)$ in Σ_+ and a constant $c > 0$ such that, for all $X \in \mathcal{V}$ and for all $Z \in \mathcal{E}^-(X)$, one has*

$$|B(X)\,Z|^2 \geq c\,\gamma^2 \left(|Z_3^r|^2 + |Z_2^r|^2\right).$$

An analogous estimate holds in a neighborhood of points $(-V_1|\eta|,\eta,0)$.

Proof. According to Lemma 3.1 we know that the kernel of the restriction of B to the stable subspace $\mathcal{E}^-$ at the point $(V_1|\eta|,\eta,0)$ is a one-dimensional space. Therefore, in order to prove Lemma 3.3, we need only to show that 0 is a *simple root* of the determinant of the restriction of B to $\mathcal{E}^-$ or, more precisely, that the partial derivative of this determinant with respect to γ calculated at $\gamma = 0$ is not zero.

We first deal with the case $d = 2$, and we keep the notation a_3^r introduced in the proof of Lemma 3.1. After a few simplifications, for $Z \in \mathcal{E}^-$, we get

$$B(\delta,\eta,\gamma)\, Z = \begin{pmatrix} \rho_r(c_r^2\tau + u_r a_3^r) & 2ij\eta \\ \dfrac{ij\eta\widetilde{\gamma}(c_r^2\tau + u_l a_3^r)}{\sqrt{\gamma^2+\delta^2+\widetilde{V}^2\eta^2}} & \dfrac{-\rho_r\widetilde{\gamma}(\tau^2 + u_r u_l \eta^2)}{\sqrt{\gamma^2+\delta^2+\widetilde{V}^2\eta^2}} \end{pmatrix} \begin{pmatrix} Z_3^r \\ Z_2^r \end{pmatrix}.$$

Note that this expression involves $\widetilde{\gamma}$ and some square roots because of the homogeneity property of the mapping Q. The determinant of the restriction of B to the stable subspace $\mathcal{E}^-$ is therefore given by

$$\det B^- = \frac{i\widetilde{\gamma}\rho_r^2|\eta|^3}{\sqrt{\gamma^2+\delta^2+\widetilde{V}^2\eta^2}} \underbrace{\left[c_r^2 V(V^2+2u_r^2-u_r u_l) + u_r A_3^r (V^2+u_r u_l)\right]}_{f(V)},$$

where V and A_3^r denote the same reduced quantities as those defined in the proof of Lemma 3.1. One can check that $f(V)$ vanishes at the points where the uniform stability condition fails (thanks to the expression of the Lopatinskii determinant). The final step consists of calculating the partial derivative of $\det B^-$ with respect to γ at $\gamma = 0$. Proving that this derivative is not zero is equivalent to proving that the derivative (with respect to V) of the function $f(V)$ calculated at $V = \pm V_1$ is not zero. We have

$$f'(V) = c_r(3V^2 + 2u_r^2 - u_r u_l) + \frac{c_r u_r V\left[3V^2 + u_r u_l - 2(c_r^2 - u_r^2)\right]}{u_r V A_3^r},$$

and thus, using the expression of VA_3^r at $(V_1|\eta|,\eta,0)$, we find the expression

$$f'(V_1) = c_r^2(3V_1^2 + 2u_r^2 - u_r u_l) - \frac{2c_r^2 u_r^2 V_1^2\left[3V_1^2 + u_r u_l - 2(c_r^2 - u_r^2)\right]}{(c_r^2+u_r^2)V_1^2 - (c_r^2-u_r^2)u_r u_l}.$$

Eventually, $f'(V_1) = 0$ if and only if V_1^2 is a root of the polynomial

$$Q_1(X) = 3(c_r^2-u_r^2)X^2 + 2\left[u_r^2(3c_r^2-u_r^2) - 2u_r u_l c_r^2\right]X + u_r u_l(c_r^2-u_r^2)(u_r u_l - 2u_r^2).$$

Assume that $Q_1(V_1^2) = 0$. Since V_1^2 is also a root of the polynomial P_1 defined in Lemma 3.1, we get the relation

$$\left[u_r u_l(c_r^2+3u_r^2) - u_r^2(3c_r^2+u_r^2)\right]V_1^2 - u_r u_l(c_r^2-u_r^2)(u_r u_l + u_r^2) = 0,$$

and one easily checks that the previous term between brackets is positive, since $u_r u_l > c_r^2 + u_r^2$. Plugging this explicit expression of V_1^2 into the definition of P_1 implies that $S := u_r u_l / c_r^2$ is a root of the following polynomial:

$$Q_2(X) = (1-M_r^2)X^3 + (2M_r^4 + 3M_r^2 - 1)X^2 - M_r^2(M_r^4 + 5M_r^2 + 2)X + M_r^4(3+M_r^2).$$

One easily checks that $Q_2(1) = 0$, and we have assumed that $S > 1 + M_r^2$. We thus deduce that S is a root of the polynomial

$$Q_3(X) = (1-M_r^2)X^2 + 2M_r^2(1+M_r^2)X - M_r^4(3+M_r^2) = 0.$$

However, the value of $Q_3(1+M_r^2)$ is greater than 1, so S is always larger than the greatest root of Q_3. We are thus led to a contradiction. Therefore V_1^2 cannot be a root of the polynomial Q_1 which means exactly that $f'(V_1) \neq 0$.

If $d = 3$ and Z is a vector in the stable subspace $\mathcal{E}^-$, we have the relation

$$B(\delta,\eta,\gamma)\,Z = \begin{pmatrix} \rho_r(c_r^2\tau + u_r a_3^r) & 2ij\eta_1 & 2ij\eta_2 \\ \dfrac{ij\eta_1\widetilde{\gamma}(c_r^2\tau + u_l a_3^r)}{\sqrt{\gamma^2+\delta^2+\widetilde{V}^2|\eta|^2}} & \dfrac{-\rho_r\widetilde{\gamma}(\tau^2 + u_r u_l \eta_1^2)}{\sqrt{\gamma^2+\delta^2+\widetilde{V}^2|\eta|^2}} & \dfrac{-\rho_r\widetilde{\gamma}u_r u_l \eta_1\eta_2}{\sqrt{\gamma^2+\delta^2+\widetilde{V}^2|\eta|^2}} \\ \dfrac{ij\eta_2\widetilde{\gamma}(c_r^2\tau + u_l a_3^r)}{\sqrt{\gamma^2+\delta^2+\widetilde{V}^2|\eta|^2}} & \dfrac{-\rho_r\widetilde{\gamma}u_r u_l \eta_1\eta_2}{\sqrt{\gamma^2+\delta^2+\widetilde{V}^2|\eta|^2}} & \dfrac{-\rho_r\widetilde{\gamma}(\tau^2 + u_r u_l \eta_1^2)}{\sqrt{\gamma^2+\delta^2+\widetilde{V}^2|\eta|^2}} \end{pmatrix} \begin{pmatrix} Z_3^r \\ \\ Z_2^r \end{pmatrix},$$

from which we get the expression

$$\det B^- = \frac{i\widetilde{\gamma}^2\rho_r^3 V^2|\eta|^5}{\gamma^2+\delta^2+\widetilde{V}^2\eta^2}\, f(V)\,.$$

Therefore the previous analysis made in the case $d = 2$ applies, and the conclusion of the lemma follows. □

In order to simplify what follows, we assume that the reference speed $\widetilde{V}$ and the reference frequency $\widetilde{\gamma}$ are *normalized* and taken to be equal to 1. This is of pure convenience and does not affect the following results, but it will clarify the introduction of weighted Sobolev spaces.

3.2. A priori estimate on the linearized equations. We begin with a result of existence of a microlocal Kreiss symmetrizer for system (2.8). The proof of this result is detailed in the next subsection. Except at the particular points where the uniform stability condition fails, the method is the one developed in [20] (see also [8]) whose first purpose was the resolution of mixed initial boundary value problems for strictly hyperbolic systems when the boundary conditions do not have any "dissipativeness" property. We point out that this method was later used in [25] (see also [27, 30]) to deal with multidimensional shock waves where no "dissipativeness" argument holds, since the boundary conditions B take the form of a pseudodifferential operator of order 0. In our case, since we have limited the study to constant coefficients systems, these boundary conditions take the simpler form of a Fourier multiplier.

We shall see in the proof of Theorem 3.5 that the failure of the uniform stability condition in the so-called hyperbolic region gives rise to some poor energy estimates compared to the maximal L^2 estimates obtained under the uniform stability condition. In fact, we can state the following result.

PROPOSITION 3.4. *For all $X_0 \in \Sigma_+$, there exists an open neighborhood $\mathcal{V}$ of X_0 and matrices $r(X)$, $T(X)$ of class C^∞ with respect to $X \in \mathcal{V}$ which satisfy the following:*

$r(X)$ is hermitian.

$T(X)$ is invertible, and, defining $a(X) = T(X)^{-1}\mathcal{A}(X)\,T(X)$, $\widetilde{B}(X) = B(X)\,T(X)$, there exist two positive constants C and $c > 0$ such that

$$\operatorname{Re}\,(r(X)\,a(X)) \geq c\gamma I\,,$$
$$r(X) + C\widetilde{B}(X)^*\widetilde{B}(X) \geq cI\,,$$

if the Lopatinskii determinant does not vanish at X_0, and

$$\text{Re } (r(X)\, a(X)) \geq c\gamma^3 I\,,$$
$$r(X) + C\widetilde{B}(X)^*\widetilde{B}(X) \geq c\gamma^2 I\,,$$

if X_0 is a root of the Lopatinskii determinant. In this latter case, $r(X)$ can be chosen under the following diagonal form:

$$r(X) = \begin{pmatrix} -\gamma^2 I_d & 0 \\ 0 & \lambda I_{d+2} \end{pmatrix},$$

where λ is a real number greater than 1.

Recall that, under the uniform stability condition, one can construct a Kreiss symmetrizer R that satisfies

$$\text{Re } (R(X)\mathcal{A}(X)) \geq c\gamma I\,,$$
$$R(X) + CB(X)^*B(X) \geq cI\,.$$

Proposition 3.4 enables us to derive an energy estimate on system (2.8) in some appropriate weighted spaces. We define two domains Ω and ω as

$$\Omega = \mathbb{R}\times\mathbb{R}^d_+ = \{(t,y,z)\in\mathbb{R}^{d+1} \text{ s.t. } z>0\} \quad \text{and} \quad \omega = \mathbb{R}\times\mathbb{R}^{d-1} = \partial\overline{\Omega}\,.$$

For $\gamma>0$ and $s\in\mathbb{R}$ we define the following symbols:

$$\forall \xi\in\mathbb{R}^d, \quad \lambda^{s,\gamma}(\xi) = (\gamma^2+|\xi|^2)^{s/2}\,.$$

The usual Sobolev spaces $H^s(\omega)$ are equipped with the following weighted norms (depending on the positive parameter γ):

$$\|u\|^2_{s,\gamma} = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \lambda^{2s,\gamma}(\xi)\,|\widehat{u}(\xi)|^2\,d\xi\,.$$

These weighted norms enable us to construct a parameter version of the classical pseudodifferential calculus which is of constant use in the study of mixed initial boundary value problems for hyperbolic systems; see [1, 20, 25].

For all integer k, we equip the usual Sobolev space $H^k(\Omega)$ with the following norm:

$$|\!|\!| U |\!|\!|^2_{k,\gamma} = \sum_{j=0}^{k}\int_0^{+\infty} \|\partial_z^j U(.,z)\|^2_{k-j,\gamma}\,dz\,.$$

We now define two operators $\mathcal{L}$ and $\mathcal{B}$ by

$$\mathcal{L}(U) = \partial_t U + \sum_{j=1}^{d-1}\mathcal{A}_j\,\partial_{x_j}U + \mathcal{A}_d\,\partial_z U \quad \text{for } z>0,$$

$$\mathcal{B}(\varphi,U) = \partial_t\varphi\, b_0 + \sum_{j=1}^{d-1}\partial_{x_j}\varphi\, b_j + M\,U \quad \text{for } z=0.$$

The change of unknown functions described in section 2 leads to the introduction of the "weighted" operators

$$\mathcal{L}^\gamma(U) = \mathcal{L}(U) + \gamma U \quad \text{and} \quad \mathcal{B}^\gamma(\varphi,U) = \mathcal{B}(\varphi,U) + \gamma\varphi\, b_0\,.$$

These notations enable us to state our first weak stability theorem.

THEOREM 3.5. *There exists a constant $C > 0$ such that, for all $U \in H^2(\Omega)$, for all $\varphi \in H^2(\omega)$, and for all $\gamma \geq 1$, the following estimate holds:*

$$\gamma |\!|\!| U |\!|\!|_{0,\gamma}^2 + \|U\|_{0,\gamma}^2 + \|\varphi\|_{1,\gamma}^2 \leq C \left(\frac{1}{\gamma^3} |\!|\!| \mathcal{L}^\gamma U |\!|\!|_{1,\gamma}^2 + \frac{1}{\gamma^2} \|\mathcal{B}^\gamma(\varphi, U)\|_{1,\gamma}^2 \right).$$

We recall that, under the uniform stability condition, one deduces from the existence of a global Kreiss symmetrizer the following maximal L^2 estimate:

$$\gamma |\!|\!| U |\!|\!|_{0,\gamma}^2 + \|U\|_{0,\gamma}^2 + \|\varphi\|_{1,\gamma}^2 \leq C \left(\frac{1}{\gamma} |\!|\!| \mathcal{L}^\gamma U |\!|\!|_{0,\gamma}^2 + \|\mathcal{B}^\gamma(\varphi, U)\|_{1,\gamma}^2 \right).$$

Comparing to the result of Theorem 3.5, we see that losses of derivatives appear both in the interior domain and on the boundary. This is quite a remarkable difference between our study and previous works such as [11, 34], where derivatives were lost only on the boundary.

Proof. The result is a consequence of the existence of a symbolic symmetrizer r given by Proposition 3.4. Since Σ_+ is a compact set, we can fix a finite covering $(\mathcal{V}_i)_{1 \leq i \leq I}$ of Σ_+ by open sets defined in Proposition 3.4. Let $(\psi_i)_{1 \leq i \leq I}$ be a partition of unity associated with this covering. More precisely, the functions ψ_i are nonnegative, C^∞, and satisfy

$$\forall i = 1, \dots, I, \quad \text{Supp } \psi_i \subset \mathcal{V}_i \quad \text{and} \quad \sum_{i=1}^{I} \psi_i^2 \equiv 1 .$$

Now let $U \in H^2(\Omega)$ and $\varphi \in H^2(\omega)$. We denote $\widehat{U}(\xi, z)$ the Fourier transform of $U(t, y, z)$ with respect to the d first variables (t, y). We also define

$$F(t, y, z) = \mathcal{L}^\gamma U(t, y, z) \in H^1(\Omega) ,$$
$$G(t, y) = \mathcal{B}^\gamma(\varphi, U) \in H^1(\omega) .$$

Lemma 3.2 ensures that there exists a constant $C > 0$ such that

$$\lambda^{2,\gamma}(\xi) \, |\widehat{\varphi}(\xi)|^2 \leq C \left(|\widehat{U}(\xi, 0)|^2 + |\widehat{G}(\xi)|^2 \right) ,$$

with $\xi = (\delta, \eta)$. Integrating with respect to ξ and using Plancherel's theorem yield the inequalities

$$\begin{aligned} \|\varphi\|_{1,\gamma}^2 &\leq C \left(\|U\|_{0,\gamma}^2 + \|\mathcal{B}^\gamma(\varphi, U)\|_{0,\gamma}^2 \right) \\ &\leq C \left(\|U\|_{0,\gamma}^2 + \gamma^{-2} \|\mathcal{B}^\gamma(\varphi, U)\|_{1,\gamma}^2 \right) . \end{aligned}$$

We now need to estimate the norms $\|U\|_{0,\gamma}^2$ and $|\!|\!| U |\!|\!|_{0,\gamma}^2$ in terms of $\|G\|_{1,\gamma}^2$ and $|\!|\!| F |\!|\!|_{1,\gamma}^2$. We define

$$V_i(X, z) = \psi_i(X) \, T_i(X)^{-1} \, \widehat{U}(\xi, z) .$$

Since ψ_i has compact support in $\mathcal{V}_i$, we extend the mappings r_i and T_i on all Σ_+, assuming them to be constant outside of $\mathcal{V}_i$. (This is of pure convenience since only the value of these mappings on Supp ψ_i will be involved in what follows.) Then we extend r_i and T_i (and thus a) as homogeneous functions of degree 0 in $X = (\xi, \gamma)$. (This is the method developed in [8, 20, 30].)

Using the definition of the matrix $a(X)$, we know that $V_i(X,z)$ satisfies the ordinary differential equation

$$\frac{dV_i}{dz} = a(X)\,V_i + \psi_i(X)\,T_i(X)^{-1}\,\mathcal{A}_d^{-1}\,\widehat{F}\,.$$

We first deal with the case where $\mathcal{V}_i$ is a neighborhood of a root of the Lopatinskii determinant. We take the scalar product of the previous equation by $\lambda^{2,\gamma}(\xi)\,r_i(X)V_i$ and integrate with respect to $\xi=(\delta,\eta)\in\mathbb{R}^d$. Then we integrate with respect to z from 0 to $+\infty$. Using the properties of the symmetrizer r_i, we get

$$\begin{aligned}-2\ \mathrm{Re}\ \left\langle\left\langle r_i(X)\,V_i, \psi_i(X)\,\lambda^{2,\gamma}(\xi)\,T_i(X)^{-1}\,\mathcal{A}_d^{-1}\widehat{F}\right\rangle\right\rangle&\\ \geq c\gamma^2\left\|\psi_i\widehat{U}\right\|_{0,\gamma}^2 - C\left\|\psi_i B\,\widehat{U}\right\|_{1,\gamma}^2 + 2\ \mathrm{Re}\ \langle\langle V_i, \lambda^{2,\gamma}\,r_i(X)\,a(X)\,V_i\rangle\rangle\,.&\end{aligned}$$

Define a matrix Σ as

$$\Sigma = \begin{pmatrix} \dfrac{\gamma}{\sqrt{\gamma^2+|\xi|^2}} & \mathbf{0} \\ \mathbf{0} & \sqrt{\lambda} \end{pmatrix},$$

where λ is a real number greater than 1 as stated in Proposition 3.4. We clearly have $\mathrm{Re}\ r_i(X)\,a(X)\geq c\gamma\Sigma^2$ for X in the support of ψ_i. Since a and r_i are diagonal matrices on $\mathcal{V}_i$, we have

$$2\ \mathrm{Re}\ \langle\langle V_i, \lambda^{2,\gamma}\,r_i(X)\,a(X)\,V_i\rangle\rangle \geq c\gamma|\!|\!|\lambda^{1,\gamma}\,\Sigma V_i|\!|\!|_{0,\gamma}^2\,,$$

and the Cauchy–Schwarz inequality yields the estimate

$$\begin{aligned}-2\ \mathrm{Re}\ \left\langle\left\langle r_i(X)\,V_i, \psi_i(X)\,\lambda^{2,\gamma}(\xi)\,T_i(X)^{-1}\,\mathcal{A}_d^{-1}\widehat{F}\right\rangle\right\rangle &\leq c\gamma|\!|\!|\lambda^{1,\gamma}\,\Sigma V_i|\!|\!|_{0,\gamma}^2 + \frac{C}{\gamma}\left|\!\left|\!\left|\lambda^{1,\gamma}\,\Sigma\widehat{F}\right|\!\right|\!\right|_{0,\gamma}^2\\ &\leq c\gamma|\!|\!|\lambda^{1,\gamma}\,\Sigma V_i|\!|\!|_{0,\gamma}^2 + +\frac{C}{\gamma}|\!|\!|F|\!|\!|_{1,\gamma}^2\,.\end{aligned}$$

Eventually, we get the following estimate:

$$c\gamma^2\left\|\psi_i\widehat{U}\right\|_{0,\gamma}^2 + c\gamma|\!|\!|\lambda^{1,\gamma}\,\Sigma V_i|\!|\!|_{0,\gamma}^2 \leq \frac{C}{\gamma}|\!|\!|F|\!|\!|_{1,\gamma}^2 + C\left\|\psi_i B\,\widehat{U}\right\|_{1,\gamma}^2\,,$$

from which we finally obtain

$$\gamma^2\left\|\psi_i\widehat{U}\right\|_{0,\gamma}^2 + \gamma^3\left\|\psi_i\widehat{U}\right\|_{0,\gamma}^2 \leq \frac{C}{\gamma}|\!|\!|F|\!|\!|_{1,\gamma}^2 + C\left\|\psi_i B\,\widehat{U}\right\|_{1,\gamma}^2\,.$$

When $\mathcal{V}_i$ is a neighborhood of a point X_0, where the Lopatinskii determinant does not vanish, the result is directly obtained by the analysis made by Kreiss [20] (see also [8, 30]) which gives the maximal L^2 estimate. All these inequalities give an estimate on U in terms of $\mathcal{L}^\gamma(U)$ and $\mathcal{B}^\gamma(\varphi,U)$. The previous estimate on the front φ added to this estimate on U gives the result. □

Note that, when $\mathcal{L}^\gamma(U)=0$, we recover Majda's statement on weakly stable shocks (see [25, p. 10]). However, Theorem 3.5 is a little more precise, since it indicates two types of loss of derivatives arising in this problem. Some regularity is lost on the boundary, as pointed out in Majda's work. However, in addition, a very severe loss of regularity occurs in the domain Ω.

3.3. Construction of a Kreiss symmetrizer: Proof of Proposition 3.4. In this subsection, we prove Proposition 3.4 and construct a microlocal symmetrizer. This construction relies on the so-called block structure of the symbol $\mathcal{A}$ which was introduced by Kreiss in the case of strictly hyperbolic systems [20]. In [25], Majda extended this property in a general definition and proved that isentropic Euler equations (2.1) met all the requirements. We point out that, in a recent paper [28], Métivier succeeded in proving that Majda's definition of the block structure condition was a property satisfied by all hyperbolic systems of conservation laws with constant multiplicity eigenvalues.

We need to distinguish four cases corresponding to the different behaviors of the eigenmodes $\omega_k^{l,r}$. We recall that, when $\gamma = 0$, the eigenmodes ω_1^l and ω_3^l are always distinct (see section 2).

Construction of r in the elliptic region. Let $X_0 \in \Sigma_+$ such that $\gamma > 0$. The symbol $\mathcal{A}(X_0)$ has no purely imaginary eigenvalue, and one can therefore choose two closed curves C^- (resp., C^+) lying in the half-plane $\{\text{Re } z < 0\}$ (resp., $\{\text{Re } z > 0\}$) such that the eigenvalues of negative (resp., positive) real part of $\mathcal{A}(X_0)$ stand in the domain delimited by C^- (resp., C^+). Using the generalized eigenprojectors associated with $C^\pm$, one gets the existence of a C^∞ mapping $T(X)$ with values in the set of $2(d+1) \times 2(d+1)$ invertible matrices, defined on a neighborhood of X_0, such that

$$\forall X \in \mathcal{V}\,, \quad T(X)^{-1}\mathcal{A}(X)\,T(X) = \begin{pmatrix} a^-(X) & \mathbf{0} \\ \mathbf{0} & a^+(X) \end{pmatrix},$$

and the spectrum of $a^-(X)$ (resp., $a^+(X)$) is contained in the half-space $\{\text{Re } z < 0\}$ (resp., $\{\text{Re } z > 0\}$).

Now define the positive definite hermitian matrices

$$H^- = 2\int_0^{+\infty} \exp(ta^-(X_0))^* \exp(ta^-(X_0))\,dt$$

and

$$H^+ = 2\int_0^{+\infty} \exp(-ta^+(X_0))^* \exp(-ta^+(X_0))\,dt\,.$$

One easily checks that

$$\begin{aligned} \text{Re } (H^+\,a^+(X_0)) &:= (H^+\,a^+(X_0) + a^+(X_0)^* H^+)/2 = I\,, \\ \text{Re } (H^-\,a^-(X_0)) &:= (H^-\,a^-(X_0) + a^-(X_0)^* H^-)/2 = -I\,. \end{aligned}$$

This is the classical Lyapunov matrix theorem; see [2]. In a neighborhood $\mathcal{V}$ of X_0, one has

$$\forall\, X \in \mathcal{V}\,, \quad \text{Re } H^- a^-(X) \le -\frac{1}{2} I \quad \text{and} \quad \text{Re } H^+ a^+(X) \ge \frac{1}{2} I\,.$$

We now define

$$r = \begin{pmatrix} -H^- & 0 \\ 0 & \lambda H^+ \end{pmatrix},$$

where λ will be a real number fixed greater than 1 in what follows. It is clear that r satisfies the first property of the lemma. Moreover, if Z denotes any vector of $\mathbb{C}^{2(d+1)}$, we can write

$$\widetilde{B}(X_0)\, Z = \widetilde{B}(X_0) \begin{pmatrix} Z^- \\ 0 \end{pmatrix} + \widetilde{B}(X_0) \begin{pmatrix} 0 \\ Z^+ \end{pmatrix} .$$

Since the Lopatinskii determinant does not vanish at any point of $\mathcal{V}$, there exists a constant $C > 0$ such that

$$|Z^-|^2 \leq C \left(|Z^+|^2 + |\widetilde{B}(X_0)\, Z|^2 \right) .$$

Following [8, 20], one can check that, for sufficiently large λ, we have

$$r + C\widetilde{B}(X_0)^* \widetilde{B}(X_0) \geq cI\,,$$

for some constant $c > 0$, and this estimate holds in all $\mathcal{V}$ by a continuity argument (replacing c by $c/2$).

Construction of r at a hyperbolic diagonalization point. Let $X_0 \in \Sigma_+$ such that $\gamma = 0$, $\eta \neq 0$, and $\delta \neq \pm|\eta|\sqrt{c_{r,l}^2 - u_{r,l}^2}$. We also assume that the Lopatinskii determinant does not vanish at X_0 and therefore does not vanish in a suitable neighborhood of X_0. Using the parametrization of the eigenspaces associated with the eigenmodes $\omega^{l,r}$, it is clear that one can construct a C^∞ mapping T such that, for all X in a neighborhood $\mathcal{V}$ of X_0, one has

$$\forall X \in \mathcal{V}\,, \quad T(X)^{-1} \mathcal{A}(X)\, T(X) = \begin{pmatrix} \omega_3^r & & & & & \\ & \omega_2^r I_{d-1} & & & \mathbf{0} & \\ & & \omega_1^r & & & \\ & & & \omega_1^l & & \\ & \mathbf{0} & & & \omega_2^l I_{d-1} & \\ & & & & & \omega_3^l \end{pmatrix} .$$

To achieve the construction of the symmetrizer in this case, we first need to study the behavior of ω_1^r and ω_3^r near X_0. We shall prove in section 5 that there exists a constant $c > 0$ such that

$$\forall\, X \in \mathcal{V}\,, \quad \begin{cases} -\mathrm{Re}\ \omega_3^r \geq c\gamma\,, \\ \ \ \mathrm{Re}\ \omega_1^r \geq c\gamma\,. \end{cases}$$

Similar results hold for the behavior of the eigenmodes ω_1^l and ω_3^l. Then it is sufficient to choose r under diagonal form

$$r = \begin{pmatrix} -1 & & & & & \\ & -I_{d-1} & & & \mathbf{0} & \\ & & \lambda & & & \\ & & & \lambda & & \\ & \mathbf{0} & & & \lambda I_{d-1} & \\ & & & & & \lambda \end{pmatrix} ,$$

and performing the same analysis as in the elliptic region yields the required properties on the symmetrizer r.

Construction of r in the neighborhood of Jordan points. Let $X_0 \in \Sigma_+$ such that $\gamma = 0$ and $\delta = \pm|\eta|\sqrt{c_r^2 - u_r^2}$. Using the same type of arguments as in the case $\gamma > 0$, one can prove that there exists a C^∞ mapping $T(X)$ with values in the set of $2(d+1) \times 2(d+1)$ invertible matrices, defined on a neighborhood of X_0, such that

$$\forall X \in \mathcal{V}\,, \quad T(X)^{-1}\mathcal{A}(X)\,T(X) = \begin{pmatrix} \omega_2^r I_{d-1} & & & & \\ & a_r(X) & & \mathbf{0} & \\ & & \omega_1^l & & \\ & \mathbf{0} & & \omega_2^l I_{d-1} & \\ & & & & \omega_3^l \end{pmatrix},$$

with $a_r(X)$ some 2×2 matrix satisfying

$$a_r(X_0) = \begin{pmatrix} \lambda_r & i \\ 0 & \lambda_r \end{pmatrix},$$

$\lambda_r = i\kappa_r$ being the double (purely imaginary) root of the polynomial

$$(c_r^2 - u_r^2)X^2 \pm 2i|\eta|u_r\sqrt{c_r^2 - u_r^2}X - u_r^2|\eta|^2\,,$$

which is nothing but (2.10) at point X_0. We shall show in section 5 that T can be chosen such that, for all $X \in \mathcal{V} \cap \{\gamma = 0\}$, $a_r(X)$ has purely imaginary coefficients. Furthermore, if $D_r(X)$ denotes the partial derivative of $a_r(X)$ with respect to γ, the lower left corner coefficient α_r of $D_r(X_0)$ is a nonzero real number.

We define $r(X)$ in the following way:

$$r(X) = \begin{pmatrix} -1 & & & & \\ & -I_{d-1} & & \mathbf{0} & \\ & & h_r(X) & & \\ & \mathbf{0} & & \lambda I_{d-1} & \\ & & & & \lambda \end{pmatrix},$$

λ once again being some real number greater than 1 fixed in what follows. Following the analysis of Kreiss [8, 20], we choose h_r of the form

$$h_r(X) = \underbrace{\begin{pmatrix} 0 & e_1 \\ e_1 & e_2 \end{pmatrix}}_{E} + \underbrace{\begin{pmatrix} f(X) & 0 \\ 0 & 0 \end{pmatrix}}_{F(X)} - i\gamma \underbrace{\begin{pmatrix} 0 & -g \\ g & 0 \end{pmatrix}}_{G},$$

where e_1, e_2, and g are real numbers and f is a C^∞ real-valued function that we shall fix in what follows. The Taylor expansion of $a_r(X)$ reads

$$a_r(X) = i\left(\kappa_r I + N - iB_r\left(\widetilde{X}\right)\right) + \gamma D_r\left(\widetilde{X}\right) + \gamma^2 M(X)\,,$$

where $\widetilde{X} = (\delta, \eta, 0)$ if $X = (\delta, \eta, \gamma)$, and $B_r(\widetilde{X}) = a_r(\widetilde{X}) - a_r(X_0)$; in the previous relation, N denotes the nilpotent matrix

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

We know that B_r reads

$$B_r\left(\widetilde{X}\right) = i\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix},$$

with real-valued C^∞ functions b_{ij} vanishing at X_0. We fix f by the following formula:

$$f(X) = \frac{e_1(b_{11} - b_{22}) + e_2 b_{21}}{1 + b_{12}}$$

so that f has the required property. Moreover, this choice of f implies that

$$(E + F(X))\left(N - iB_r\left(\widetilde{X}\right)\right)$$

is a real symmetric matrix. As a consequence, one gets

$$\text{Re}\,(h_r(X)\,a_r(X)) = \gamma\ \text{Re}\,\left(G\,N\,+\,E\,D_r\left(\widetilde{X}\right)\right) + \gamma\,L(X)\,,$$

where L is a C^∞ hermitian matrix which vanishes at X_0. The shape of E and G yields

$$\text{Re}\,(G\,N\,+\,E\,D_r(X_0)) = \begin{pmatrix} 0 & 0 \\ 0 & g \end{pmatrix} + \begin{pmatrix} e_1\alpha_r & * \\ * & * \end{pmatrix},$$

where quantities denoted by $*$ depend only on e_1 and e_2. We fix $e_1 = 1/\alpha_r$ and g sufficiently large so that

$$\text{Re}\,(h_r(X)\,a_r(X)) \geq c\gamma I\,.$$

This is possible as long as the choice of e_2 does not depend on g. In fact, e_2 will be fixed in order to give the estimate with respect to the boundary conditions $\widetilde{B}$, and the choice will not involve g. Indeed, the choice of h_r implies that

$$r(X_0) = \begin{pmatrix} -I_{d-1} & & & & & \\ & 0 & e_1 & & \mathbf{0} & \\ & e_1 & e_2 & & & \\ & & & \lambda & & \\ & \mathbf{0} & & & \lambda I_{d-1} & \\ & & & & & \lambda \end{pmatrix},$$

and a rather tedious analysis (essentially based on the Cayley–Hamilton theorem) shows that the stable subspace $\mathcal{E}^-(X_0)$ is spanned by the d first vectors of our new basis. Since the Lopatinskii determinant does not vanish in $\mathcal{V}$, we can therefore fix sufficiently large e_2 and λ (independently of g) to get an estimate of the type

$$r + C\widetilde{B}(X_0)^*\widetilde{B}(X_0) \geq cI\,.$$

An appropriate choice of g achieves the construction.

We now turn to the last case of points where the uniform stability condition fails. Note that the previous result on the behavior of the eigenmodes still hold because of

the properties of V_1^2. Indeed, one can diagonalize the symbol $\mathcal{A}$ in a neighborhood of $(V_1|\eta|, \eta, 0)$; in other words, we still have the existence of a C^∞ mapping T satisfying

$$\forall X \in \mathcal{V}\,, \quad T(X)^{-1}\mathcal{A}(X)\,T(X) = \begin{pmatrix} \omega_3^r & & & & & \\ & \omega_2^r I_{d-1} & & & \mathbf{0} & \\ & & \omega_1^r & & & \\ & & & \omega_1^l & & \\ & \mathbf{0} & & & \omega_2^l I_{d-1} & \\ & & & & & \omega_3^l \end{pmatrix}.$$

To recover the estimate of r with respect to the boundary conditions B, one has to choose r of the form

$$r = \begin{pmatrix} -\gamma^2 & & & & & \\ & -\gamma^2 I_{d-1} & & & \mathbf{0} & \\ & & \lambda & & & \\ & & & \lambda & & \\ & \mathbf{0} & & & \lambda I_{d-1} & \\ & & & & & \lambda \end{pmatrix}.$$

Using Lemma 3.3 and performing the same analysis as in the elliptic region yield the estimate

$$r(X) + C\widetilde{B}(X)^*\widetilde{B}(X) \geq c\gamma^2 I$$

for sufficiently large λ. Since r is diagonal, we immediately have the estimate

$$\text{Re}\,(r\,a(X)) \geq c\gamma^3 I\,,$$

and this completes the proof of Proposition 3.4.

4. Subsonic phase transitions in a van der Waals fluid. In this section, we consider the nonstandard initial boundary value problem (2.8) with boundary conditions given by (2.7). We follow the method adopted in section 3 and begin by recalling the main result of [4].

LEMMA 4.1 (Benzoni-Gavage [4]). *There exists a positive number V_0 such that, for all $(\delta, \eta, \gamma) \in \mathbb{R}^{d+1}$ satisfying $\gamma \geq 0$ and $(\delta, \gamma) \neq (\pm iV_0|\eta|, 0)$, one has*

$$\big\{(Z, \chi) \in \mathcal{E}^-(\delta, \eta, \gamma) \times \mathbb{C} \ \text{ s.t. } \ \chi\, b(\delta, \eta, \gamma) + M\,Z = 0\big\} = \{0\}\,,$$

and, for $\eta \neq 0$, the set

$$\big\{(Z, \chi) \in \mathcal{E}^-(\pm V_0|\eta|, \eta, 0) \times \mathbb{C} \ \text{ s.t. } \ \chi\, b(\pm V_0|\eta|, \eta, 0) + M\,Z = 0\big\}$$

is a one-dimensional subspace of $\mathbb{C}^{2d+3}$. If (Z, χ) belongs to this subspace, then

$$Z_r \in \mathbb{C}\begin{pmatrix} \rho_r(\tau + u_r\,\omega_3^r) \\ -c_r^2\, i\eta \\ -c_r^2\,\omega_3^r \end{pmatrix} \quad \textit{and} \quad Z_l \in \mathbb{C}\begin{pmatrix} \rho_l(\tau - u_l\,\omega_1^l) \\ -c_l^2\, i\eta \\ c_l^2\,\omega_1^l \end{pmatrix};$$

that is, Z_r has no component on the eigenspace associated with the eigenvalue ω_2^r. At all points of the form $(\pm V_0|\eta|, \eta, 0)$, both eigenmodes ω_3^r and ω_1^l have negative real part (which explains the designation "surface waves").

By definition, V_0^2 is the positive root of the polynomial

$$P_2(X) = \frac{c_r^2 c_l^2 - u_r^2 u_l^2}{u_r^2 u_l^2} X^2 + (c_r^2 - u_r^2 + c_l^2 - u_l^2) X - (c_r^2 - u_r^2)(c_l^2 - u_l^2)\,,$$

and the following inequalities hold:

$$V_0^2 < \min(c_r^2 - u_r^2, c_l^2 - u_l^2) \quad \textit{and} \quad V_0^2 < u_r u_l\,.$$

4.1. Elimination of the front. As we did in section 3 we begin by isolating the shock front in the last boundary condition of (2.9). This is stated as follows.

LEMMA 4.2. *There exists a C^∞ mapping Q defined on the half-space $\mathbb{R}^d \times \mathbb{R}^+ \setminus \{0\}$, homogeneous of degree 0, with values in the set of square $(d+2) \times (d+2)$ invertible matrices such that, for all $X \in \mathbb{R}^d \times \mathbb{R}^+ \setminus \{0\}$, the first $d+1$ components of the vector $Q(X)\, b(X)$ vanish.*

Proof. The Rankine–Hugoniot jump relations together with (2.7) yield the relations

$$b(\delta,\eta,\gamma) = \begin{pmatrix} -\tau[\rho] \\ ij[u]\eta \\ 0 \\ -\tau[u] \end{pmatrix} \text{ if } d=2 \quad \text{and} \quad b(\delta,\eta,\gamma) = \begin{pmatrix} -\tau[\rho] \\ ij[u]\eta_1 \\ ij[u]\eta_2 \\ 0 \\ -\tau[u] \end{pmatrix} \text{ if } d=3.$$

The mapping Q is first defined on the hemisphere Σ_+ and then extended by homogeneity. Note that we go back to the first definition of Σ_+ with a reference velocity $\widetilde{V}$ and a reference frequency $\widetilde{\gamma}$ to take the physical dimension of the quantities into account.

One easily checks that, for $d=2$, the matrix

$$Q(\delta,\eta,\gamma) = \begin{pmatrix} [u] & 0 & 0 & -[\rho] \\ 0 & \tau & 0 & ij\eta \\ 0 & 0 & 1 & 0 \\ 0 & i\widetilde{V}^2\eta & 0 & j\overline{\tau} \end{pmatrix}$$

satisfies all required properties. For $d=3$, one can choose, for instance,

$$Q(\delta,\eta,\gamma) = \begin{pmatrix} [u] & 0 & 0 & 0 & -[\rho] \\ 0 & \tau & 0 & -i\widetilde{V}\eta_2 & ij\eta_1 \\ 0 & 0 & \tau & i\widetilde{V}\eta_1 & ij\eta_2 \\ 0 & -i\widetilde{V}\eta_2 & i\widetilde{V}\eta_1 & \overline{\tau} & 0 \\ 0 & i\widetilde{V}^2\eta_1 & i\widetilde{V}^2\eta_2 & 0 & j\overline{\tau} \end{pmatrix}$$

which also satisfies all required properties. This completes the proof. ∎

We can therefore write boundary conditions for the linearized problem (2.8) in the equivalent way

$$\begin{pmatrix} B(\delta,\eta,\gamma) \\ \ell(\delta,\eta,\gamma) \end{pmatrix} V(0) + \chi \begin{pmatrix} \mathbf{0}_{d+1} \\ \beta(\delta,\eta,\gamma) \end{pmatrix} = Q(\delta,\eta,\gamma)\, G\,,$$

where $\beta(\delta,\eta,\gamma)$ is given by

$$\beta(\delta,\eta,\gamma) = -j[u]\widetilde{\gamma}\sqrt{\gamma^2 + \delta^2 + \widetilde{V}^2|\eta|^2} \neq 0\,,$$

and this relation holds for $d = 2$ and $d = 3$. We now turn to the study of the behavior of the restriction of $B(\delta, \eta, \gamma)$ to the stable subspace $\mathcal{E}^-$ in the neighborhood of the points where the uniform stability condition fails. According to Lemma 4.1, the symbol $\mathcal{A}$ is diagonalizable in the neighborhood of such points.

We decompose all vectors Z belonging to the stable subspace $\mathcal{E}^-$ on the three different eigenspaces, denoting by Z_3^r, Z_2^r, and Z_1^- the components of Z on the eigenspaces associated with the eigenmodes ω_3^r, ω_2^r, and ω_1^l. More precisely, we decompose Z as

$$Z = \begin{pmatrix} Z_r \\ Z_l \end{pmatrix} \quad \text{with} \quad Z_r = Z_3^r \begin{pmatrix} \rho_r(\tau + u_r\,\omega_3^r) \\ -c_r^2\, i\eta \\ -c_r^2\,\omega_3^r \end{pmatrix} + \begin{pmatrix} 0 \\ -\omega_2^r Z_2^r \\ i\eta \cdot Z_2^r \end{pmatrix}$$

$$\text{and} \quad Z_l = Z_1^l \begin{pmatrix} \rho_l(\tau - u_l\,\omega_1^l) \\ -c_l^2\, i\eta \\ c_l^2\,\omega_1^l \end{pmatrix}.$$

Then we have the following microlocal estimate.

LEMMA 4.3. *There exists a neighborhood $\mathcal{V}$ of $(V_0|\eta|, \eta, 0)$ and a constant $c > 0$ such that, for all $X \in \mathcal{V}$ and for all $Z \in \mathcal{E}^-(X)$, one has*

$$|B(X)\,Z|^2 \geq c\,\gamma^2 \left(|Z_3^r|^2 + |Z_1^l|^2\right) + c|Z_2^r|^2\,.$$

An analogous estimate holds in a neighborhood of points of the form $(-V_0|\eta|, \eta, 0)$.

Proof. According to Lemma 4.1 we know that the kernel of the restriction of B to the stable subspace $\mathcal{E}^-$ is a one-dimensional space whose vectors have no Z_2^r component. Therefore, in order to prove the stated result, it is again sufficient to prove that 0 is a *simple root* of the determinant of the restriction of B to $\mathcal{E}^-$.

To avoid overloading this paper, we shall detail only the different steps of the proof in the two-dimensional case. The three-dimensional case is carried out by similar arguments, but the calculations are much more complicated due to the expression of the mapping Q defined at the previous lemma which involves the complex conjugate $\overline{\tau}$ (which was not the case in section 3).

Let $d = 2$ and define (as in the proof of Lemma 3.1) the following quantities:

$$a_3^r = \tau u_r - (c_r^2 - u_r^2)\omega_3^r, \quad a_1^l = \tau u_l + (c_l^2 - u_l^2)\omega_1^l\,.$$

Keeping the definition of the complex square root $\mathcal{R}$ introduced in the proof of Lemma 3.1, we also define two quantities $W_{r,l}(V)$ as

$$W_{r,l}(V) = \mathcal{R}(V^2 - (c_{r,l}^2 - u_{r,l}^2))\,.$$

Because of the properties of V_0 (see Lemma 4.1), both expressions W_r and W_l depend analytically on V in a neighborhood of V_0, and it is shown in [4] that V_0 also satisfies

$$c_r^2 c_l^2 V_0^2 + u_r u_l W_l(V_0) W_r(V_0) = 0\,.$$

A direct calculation shows that V_0 is a simple root of the above analytical function (as mentioned in [5]).

Now let Z be any vector in the stable subspace $\mathcal{E}^-(\delta,\eta,\gamma)$ with components Z_3^r, Z_2^r, and Z_1^l on the eigenspaces associated with the eigenmodes ω_3^r, ω_2^r, and ω_1^l. We have

$$B(\delta,\eta,\gamma)\,Z = \begin{pmatrix} \rho_r[u]a_3^r - c_r^2[\rho]\tau & i(\rho_l+\rho_r)[u]\eta & c_l^2[\rho]\tau - \rho_l[u]a_1^l \\ 0 & j\widetilde{\gamma}\dfrac{\tau^2/u_r - u_r\eta^2}{\sqrt{\gamma^2+\delta^2+\widetilde{V}^2\eta^2}} & 0 \\ \rho_r(c_r^2\tau + u_r a_3^r) & 2ij\eta & -\rho_l(c_l^2\tau + u_l a_1^l) \end{pmatrix} \begin{pmatrix} Z_3^r \\ Z_2^r \\ Z_1^- \end{pmatrix},$$

from which we get the expression of the restriction of B to the stable subspace $\mathcal{E}^-$. Letting $X=(\delta,\eta,\gamma)$, one gets the expression of the determinant of the above matrix:

$$\det B^-(X) = h_2(\gamma)\left[c_r^2c_l^2V^2 + u_ru_lW_l(V)W_r(V)\right],$$

where h_2 is given by

$$h_2(\gamma) = \frac{-j\widetilde{\gamma}c_rc_l[\rho]^2|\eta|^4(V^2+u_r^2)}{u_r\sqrt{\gamma^2+V_0^2|\eta|^2+\widetilde{V}^2|\eta|^2}}.$$

With the preceding remarks, it is now a straightforward verification that the partial derivative of this determinant with respect to γ calculated at $\gamma=0$ is not zero, simply because $h_2(0)\neq 0$.

For the three-dimensional case ($d=3$), one proceeds in the same way. The expression of the determinant of the restriction B^- is

$$\det B^-(X) = h_3(\gamma)\left[c_r^2c_l^2V^2 + u_ru_lW_l(V)W_r(V)\right],$$

where h_3 is given by

$$h_3(\gamma) = \frac{j^2\widetilde{\gamma}^3c_rc_l[\rho]^2|\eta|^6\,\overline{\tau}}{(\gamma^2+V_0^2|\eta|^2+\widetilde{V}^2|\eta|^2)^{3/2}}\left[\frac{V^4}{u_r^2}+V^2-\widetilde{V}^2\left(\frac{V^2}{u_r^2}+1\right)\frac{\tau}{\overline{\tau}}\right].$$

Once again (since $h_3(0)\neq 0$) the partial derivative of the determinant with respect to γ calculated at $\gamma=0$ is not zero. ∎

4.2. A priori estimate on the linearized equations. We begin with a result of existence of a global Kreiss symmetrizer for system (2.8).

PROPOSITION 4.4. *There exist a C^∞ mapping R defined on the half-space $\mathbb{R}^d\times\mathbb{R}_+\setminus\{0\}$, homogeneous of degree 0, and two positive constants c and C such that*

$$\text{Re }(R(X)\mathcal{A}(X)) \geq \frac{c\gamma^2}{\sqrt{\gamma^2+\delta^2+|\eta|^2}},$$
$$R(X)+CB(X)^*B(X) \geq \frac{c\gamma^2}{\gamma^2+\delta^2+|\eta|^2}$$

for all $X=(\delta,\eta,\gamma)\in\mathbb{R}^d\times\mathbb{R}_+\setminus\{0\}$.

This result will be directly derived from the microlocal analysis developed in the next subsection. We simply make the following remark: as in the study of nonuniformly stable Lax shocks for isentropic Euler equations, the failure of the uniform

stability condition yields two types of losses of derivatives. Some regularity is lost in the interior domain, and some is lost on the boundary.

The previous result enables us to derive the second main result of this paper, namely the complete energy estimate on the linearized problem (2.8) in the case of subsonic phase transitions. We keep the notations introduced in subsection 3.2 for the domains Ω, for its boundary ω, and for the linearized operators $\mathcal{L}^\gamma$ and $\mathcal{B}^\gamma$.

THEOREM 4.5. *There exists a constant $C > 0$ such that, for all $U \in H^2(\Omega)$, for all $\varphi \in H^{3/2}(\omega)$, and for all $\gamma \geq 1$, the following estimate holds:*

$$\gamma^2(|\!|\!| U |\!|\!|^2_{0,\gamma} + \|U\|^2_{-1/2,\gamma} + \|\varphi\|^2_{1/2,\gamma}) \leq C\left(\frac{1}{\gamma^2}|\!|\!| \mathcal{L}^\gamma U |\!|\!|^2_{1,\gamma} + \|\mathcal{B}^\gamma(\varphi, U)\|^2_{1/2,\gamma}\right).$$

Proof. The result is a direct consequence of the existence of a symbolic symmetrizer R given by Proposition 4.4. Let $U \in H^2(\Omega)$ and $\varphi \in H^{3/2}(\omega)$. We denote $\widehat{U}(\xi, z)$ the Fourier transform of $U(t, y, z)$ with respect to the d first variables (t, y). We also define

$$F(t, y, z) = \mathcal{L}^\gamma U(t, y, z) \in H^1(\Omega)\,,$$
$$G(t, y) = \mathcal{B}^\gamma(\varphi, U) \in H^{1/2}(\omega)\,.$$

Then Lemma 4.1 ensures that there exists a constant $C_1 > 0$ such that

$$\lambda^{1,\gamma}(\xi)\,|\widehat{\varphi}(\xi)|^2 \leq C_1 \lambda^{-1,\gamma}(\xi)\left(|\widehat{U}(\xi, 0)|^2 + |\widehat{G}(\xi)|^2\right),$$

with $\xi = (\delta, \eta)$. Integrating with respect to ξ and using Plancherel's theorem yield the estimates

$$\begin{aligned}\|\varphi\|^2_{1/2,\gamma} &\leq C_1(\|U\|^2_{-1/2,\gamma} + \|\mathcal{B}^\gamma(\varphi, U)\|^2_{-1/2,\gamma}) \\ &\leq C_1(\|U\|^2_{-1/2,\gamma} + \gamma^{-2}\|\mathcal{B}^\gamma(\varphi, U)\|^2_{1/2,\gamma})\,.\end{aligned}$$

Furthermore, $\widehat{U}$ satisfies the ordinary differential equation

$$\frac{d\widehat{U}}{dz} = \mathcal{A}(\xi, \gamma)\,\widehat{U} + \mathcal{A}_d^{-1}\,\widehat{F}\,.$$

We take the scalar product of this equation by $\lambda^{1,\gamma}(\xi)\,R(\xi, \gamma)\,\widehat{U}$ and integrate with respect to $\xi = (\delta, \eta) \in \mathbb{R}^d$. Then we integrate with respect to z from 0 to $+\infty$ and take the real part of the corresponding equality. Using the properties of the symmetrizer R, we get

$$-2\ \mathrm{Re}\ \left\langle\left\langle \widehat{U}, \lambda^{1,\gamma}(\xi)\,\mathcal{A}_d^{-1}\widehat{F}\right\rangle\right\rangle \geq 2c\gamma^2|\!|\!| U |\!|\!|^2_{0,\gamma} + 2c\gamma^2\|U\|^2_{-1/2,\gamma} - C_2\|\mathcal{B}^\gamma(\varphi, U)\|^2_{1/2,\gamma}\,.$$

The Cauchy–Schwarz inequality yields the estimate

$$-2\ \mathrm{Re}\ \left\langle\left\langle \widehat{U}, \lambda^{1,\gamma}(\xi)\,\mathcal{A}_d^{-1}\widehat{F}\right\rangle\right\rangle \leq c\gamma^2|\!|\!| U |\!|\!|^2_{0,\gamma} + \frac{C_3}{\gamma^2}|\!|\!| \mathcal{L}^\gamma U |\!|\!|^2_{1,\gamma}\,.$$

This last inequality added to the previous estimate on the front φ enables us to conclude. ∎

4.3. Construction of a Kreiss symmetrizer. We first construct a microlocal symmetrizer from which we will deduce the result of Proposition 4.4.

PROPOSITION 4.6. *For all $X_0 \in \Sigma_+$, there exists an open neighborhood $\mathcal{V}$ of X_0 and matrices $r(X)$, $T(X)$ of class C^∞ with respect to $X \in \mathcal{V}$ which satisfy the following:*

$r(X)$ is hermitian,

$T(X)$ is invertible, and, defining $a(X) = T(X)^{-1}\mathcal{A}(X)\,T(X)$, $\widetilde{B}(X) = B(X)\,T(X)$, there exist two positive constants C and c such that

$$\begin{aligned} &\text{Re }(r(X)\,a(X)) \geq c\gamma^2 I\,,\\ &r(X) + C\widetilde{B}(X)^*\widetilde{B}(X) \geq c\gamma^2 I\,. \end{aligned}$$

Proof. Many steps of the proof are identical to what has been done in the case of Lax shocks, and we shall not repeat them: in the so-called elliptic region $\{\gamma > 0\}$ and at Jordan points, the construction is entirely similar. Note that the equality $c_r^2 - u_r^2 = c_l^2 - u_l^2$ is not precluded in the context of phase transitions, though it is highly unlikely. In such a case, the reduction of $\mathcal{A}$ would involve two distinct Jordan blocks, but the microlocal construction of r would be a direct extension of what has been done in the case of a single.

The only difference relies on the properties of the symbol $\mathcal{A}$ in the neighborhood of the points where the uniform stability condition fails. Let $X_0 = (\pm V_0|\eta|, \eta, 0)$ be a point where the Lopatinskii determinant vanishes. We already know that $\mathcal{A}$ is diagonalizable in a neighborhood $\mathcal{V}$ of X_0 and that $\mathcal{V}$ may be suitably chosen so that ω_3^r and ω_1^l have negative real part in $\mathcal{V}$. We thus choose r of the form

$$r(X) = \begin{pmatrix} -\gamma^2 & & & & & \\ & -I_{d-1} & & & \mathbf{0} & \\ & & -\gamma^2 & & & \\ & & & \lambda & & \\ & \mathbf{0} & & & \lambda I_{d-1} & \\ & & & & & \lambda \end{pmatrix},$$

where λ is a real number greater than 1 which will be fixed in what follows. Since there exists a C^∞ invertible matrix $T(X)$ such that

$$\forall X \in \mathcal{V}\,, \quad T(X)^{-1}\mathcal{A}(X)\,T(X) = \begin{pmatrix} \omega_3^r & & & & & \\ & \omega_2^r I_{d-1} & & & \mathbf{0} & \\ & & \omega_1^l & & & \\ & & & \omega_3^l & & \\ & \mathbf{0} & & & \omega_2^l I_{d-1} & \\ & & & & & \omega_1^r \end{pmatrix},$$

we have Re $(r(X)\,a(X)) \geq c\gamma^2 I$ for all X in $\mathcal{V}$. We now have to fix λ in order to get the estimate on the boundary conditions. For this, we let $Z \in \mathbb{C}^{2(d+1)}$ and define Z^- (resp., Z^+) as the vector formed by the $(d+1)$ first (resp., last) components of Z. Writing $Z^- = (Z_1^-, \check{Z}^-, Z_{d+1}^-)$, Lemma 4.3 ensures that there exists a constant $c > 0$ which does not depend on Z such that

$$c\gamma^2\left(|Z_1^-|^2 + |Z_{d+1}^-|^2\right) + c|\check{Z}^-|^2 \leq C\left(|Z^+|^2 + |\widetilde{B}(X)\,Z|^2\right).$$

By the same techniques as used in the construction of the symmetrizer in the elliptic region, it is clear that, for a sufficiently large λ, there exists a constant $C > 0$ such that the following estimate holds:

$$r(X) + C\widetilde{B}(X)^*\widetilde{B}(X) \geq c\gamma^2 I\,.$$

This completes the proof of Proposition 4.6.

We can now turn to the proof of Proposition 4.4, using the gluing technique developed in [8, 30]. We fix a finite covering $(\mathcal{V}_i)_{1\leq i\leq I}$ of Σ_+ by open sets defined in Proposition 4.6. Let $(\psi_i)_{1\leq i\leq I}$ be a partition of unity associated with this covering (with the same properties as stated in section 3). We define a C^∞ mapping R on Σ_+ by the following formula:

$$\forall X \in \Sigma_+, \quad R(X) = \sum_{i=1}^{I} \psi_i^2(X) \left(T_i(X)^{-1}\right)^* r_i(X)\, T_i(X)^{-1}$$

so that R has values in the set of hermitian matrices. Moreover, we have

$$\begin{aligned}
\text{Re } (R(X)\mathcal{A}(X)) &\geq c\gamma^2 \sum_{i=1}^{I} \psi_i^2(X) \left(T_i(X)^{-1}\right)^* T_i(X)^{-1}\,,\\
R(X) + CB(X)^*B(X) &\geq c\gamma^2 \sum_{i=1}^{I} \psi_i^2(X) \left(T_i(X)^{-1}\right)^* T_i(X)^{-1}
\end{aligned}$$

for some positive constants c and C. It is clear that, for all X in the compact set Σ_+, the matrix

$$\sum_{i=1}^{I} \psi_i^2(X) \left(T_i(X)^{-1}\right)^* T_i(X)^{-1}$$

is hermitian positive definite. We can therefore conclude that there exists positive constants c and C such that, for all X in Σ_+,

$$\begin{aligned}
\text{Re } (R(X)\,\mathcal{A}(X)) &\geq c\gamma^2 I\,,\\
R(X) + CB(X)^*B(X) &\geq c\gamma^2 I\,.
\end{aligned}$$

The result of Proposition 4.4 follows by extending R in a homogeneous function of degree 0 and using the homogeneity properties of symbols $\mathcal{A}$ and B. □

We point out that the result of Theorem 4.5 is not optimal in the sense that we could define new spaces to get a refined estimate, since only 1/2 of a derivative is lost in the interior domain (and only in the tangential variables). However, we have preferred to state the result in this way to make it easier to visualize. Furthermore, the proof of the theorem appears much more simple than the proof of Theorem 3.5 where attention needs to be paid to get the best result possible.

5. Some technical lemmas. In this section, we prove three results used in the proof of Propositions 3.4 and 4.6. Though our proof uses some particular properties of system (2.1), they are essentially the same as in the general case; see [8, 20, 32].

We first begin by studying the behavior of the eigenmodes ω_1^r and ω_3^r in a neighborhood of points $X_0 = (\delta, \eta, 0)$.

LEMMA 5.1. *Let $X_0 \in \Sigma_+$ such that $\gamma = 0$, $\eta \neq 0$, and $\delta \neq \pm|\eta|\sqrt{c_r^2 - u_r^2}$. There exists a neighborhood $\mathcal{V}$ of X_0 in Σ_+ and a positive constant c such that*

$$\forall\, X \in \mathcal{V}\,, \quad \begin{cases} -\mathrm{Re}\ \omega_3^r \geq c\gamma\,, \\ \mathrm{Re}\ \omega_1^r \geq c\gamma\,. \end{cases}$$

Proof. Let $X_0 = (\delta_0, \eta_0, 0)$ satisfy the assumptions of the lemma. Using the proof of Proposition 3.4, we already know that $\mathcal{A}$ is diagonalizable in a neighborhood $\mathcal{V}$ of X_0:

$$\forall X \in \mathcal{V}\,, \quad T(X)^{-1}\mathcal{A}(X)\,T(X) = \begin{pmatrix} \omega_3^r & & & & & \\ & \omega_2^r I_{d-1} & & & \mathbf{0} & \\ & & \omega_1^r & & & \\ & & & \omega_1^l & & \\ & \mathbf{0} & & & \omega_2^l I_{d-1} & \\ & & & & & \omega_3^l \end{pmatrix}.$$

If both eigenmodes ω_1^r and ω_3^r are not purely imaginary at X_0, the result comes from a simple continuity argument. We shall therefore assume that both eigenmodes are purely imaginary at X_0. We fix $\eta = \eta_0$ and define

$$Q(\delta, \gamma, \omega) = (\omega + i\omega_1^r(X))(\omega + i\omega_3^r(X))(\omega + i\omega_2^r(X))^{d-1}\,. \tag{5.1}$$

For $\tau = \gamma + i\delta$ close to $i\delta_0$, the eigenmodes ω_k^r are pairwise distinct, and the hyperbolicity of the system (2.1) shows that, for all $\xi \in \mathbb{R}$, Q is given by

$$\begin{aligned} Q(\delta, \gamma, \xi) = \alpha \Big[\delta - i\gamma + \Big(\xi u_r + c_r\sqrt{|\eta_0|^2 + \xi^2}\Big)\Big] \\ \Big[\delta - i\gamma + \Big(\xi u_r - c_r\sqrt{|\eta_0|^2 + \xi^2}\Big)\Big]\,(\delta - i\gamma + \xi u_r)^{d-1} \end{aligned}$$

for some real constant $\alpha \neq 0$. Thus, for all real ξ, we have

$$Q(\delta, 0, \xi) \in \mathbb{R} \quad \text{and} \quad \frac{\partial Q}{\partial \gamma}(\delta, 0, \xi) \in i\mathbb{R}\,. \tag{5.2}$$

Moreover, the definition of Q gives the relation

$$\frac{\partial Q}{\partial \gamma}(\delta_0, 0, -i\omega_1^r(X_0)) = i\frac{\partial \omega_1^r}{\partial \gamma}(X_0)(-i\omega_1^r(X_0) + i\omega_3^r(X_0))(-i\omega_1^r(X_0) + i\omega_2^r(X_0))^{d-1}\,,$$

from which we conclude that the partial derivative $\partial_\gamma \omega_1^r(X_0)$ is a real number. A similar result holds for ω_3^r. We are now going to prove that this partial derivative is not zero. Equation (2.10) reads

$$(c_r^2 - u_r^2)(\omega_1^r)^2 - 2\tau u_r \omega_1^r - \tau^2 - c_r^2|\eta_0|^2 = 0\,,$$

and thus differentiating with respect to γ yields the equality

$$(c_r^2 - u_r^2)2\omega_1^r \frac{\partial \omega_1^r}{\partial \gamma} - 2u_r\omega_1^r - 2\tau u_r \frac{\partial \omega_1^r}{\partial \gamma} - 2\tau = 0\,.$$

Since ω_1^r and ω_2^r are distinct, for all $X \in \mathcal{V}$, it is clear that $\partial_\gamma \omega_1^r$ does not vanish at X_0. The end of the proof relies on a simple Taylor expansion of ω_1^r at X_0, using the fact that ω_1^r is of positive real part for $\gamma > 0$. ∎

We now turn to the study of the reduced symbol in the neighborhood of Jordan points. Let $X_0 = (\delta_0, \eta_0, 0)$ be such that

$$\delta_0 = |\eta_0|\sqrt{c_r^2 - u_r^2}$$

so that, according to the proof of Proposition 3.4, we have

$$T(X)^{-1}\mathcal{A}(X)\,T(X) = \begin{pmatrix} \omega_2^r I_{d-1} & & & \\ & a_r(X) & \mathbf{0} & \\ & & \omega_1^l & \\ & \mathbf{0} & & \omega_2^l I_{d-1} \\ & & & & \omega_3^l \end{pmatrix},$$

with $a_r(X)$ some 2×2 matrix satisfying

$$a_r(X_0) = \begin{pmatrix} \lambda_r & i \\ 0 & \lambda_r \end{pmatrix} = \lambda_r I_2 + iN\,.$$

Recall that $\lambda_r = i\kappa_r$ is the double root of the polynomial

$$(c_r^2 - u_r^2)X^2 \pm 2i|\eta|u_r\sqrt{c_r^2 - u_r^2}X - u_r^2|\eta|^2\,.$$

With these notations, we have the following result.

LEMMA 5.2. *Defining $D_r(X) = \frac{\partial a_r}{\partial \gamma}(X)$ for X close to X_0, then the lower left corner coefficient α_r of $D_r(X_0)$ is a nonzero real number.*

Proof. We fix $\eta = \eta_0$ and let $\tau = \gamma + i\delta$ be close to $i\delta_0$. We define a polynomial Q by (5.1) (see the proof of Lemma 5.1) and two polynomials Q_r and $\widetilde{Q}$ by the following formulas:

$$\begin{aligned} Q_r(\delta, \gamma, \omega) &= \det\left[\omega I_2 + i a_r(\delta, \eta_0, \gamma)\right], \\ \widetilde{Q}(\delta, \gamma, \omega) &= (\omega + i\omega_2^+)^{d-1} = (\omega - i\tau/u_r)^{d-1} \end{aligned}$$

so that $Q = Q_r\widetilde{Q}$. We already know by relation (5.2) that, for all real ξ,

$$Q(\delta, 0, \xi) \in \mathbb{R} \quad \text{and} \quad \frac{\partial Q}{\partial \gamma}(\delta, 0, \xi) \in i\mathbb{R}\,.$$

It is also clear that, for $\xi \in \mathbb{R}$, one has $\widetilde{Q}(\delta, 0, \xi) \in \mathbb{R}$.

For δ close to δ_0, $Q(\delta, 0, \omega)$ seen as a polynomial in ω has real coefficients and therefore has real roots or conjugate complex roots. Moreover, $\widetilde{Q}(\delta, 0, \omega)$ has exactly one real root, so $Q_r(\delta, 0, \omega)$ has two real roots or two conjugate complex roots. Thus, for δ close to δ_0, we have

$$\forall\, \xi \in \mathbb{R}\,, \quad Q_r(\delta, 0, \xi) \in \mathbb{R}\,. \tag{5.3}$$

The definition of λ_r yields

$$\frac{\partial Q}{\partial \gamma}(\delta_0, 0, -i\lambda_r) = \widetilde{Q}(\delta_0, 0, -i\lambda_r)\,\frac{\partial Q_r}{\partial \gamma}(\delta_0, 0, -i\lambda_r)\,,$$

and since $\lambda_r \neq -i\delta_r/u_r$ we can conclude that $\partial_\gamma Q_r(\delta_0, 0, -i\lambda_r)$ is a purely imaginary number. It is clear that 0 is a simple root of the polynomial $Q(\delta_0, \cdot, -i\lambda_r)$,

and therefore the partial derivative $\partial_\gamma Q_r(\delta_0, 0, -i\lambda_r)$ is a nonzero purely imaginary number.

To complete the proof, we note that

$$ia_r(X_0) = i\lambda_r I_2 \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix},$$

from which we get

$$\frac{\partial Q_r}{\partial \gamma}(\delta_0, 0, -i\lambda_r) = i\alpha_r \in i\mathbb{R} \setminus \{0\}. \qquad \square$$

The last thing to check is the invertible matrix $T(X)$ may be chosen in such a way that $a_r(X)$ has purely imaginary coefficients for $X \in \mathcal{V} \cap \{\gamma = 0\}$. We base our proof of this result on a technique developed in [32]. Let X_0 be the triple $(|\eta_0|\sqrt{c_r^2 - u_r^2}, \eta_0, 0)$. For $X = (\delta, \eta, \gamma)$ close to X_0, we define $\widetilde{X} = (\delta, \eta, 0)$. With these notations, we have the following result.

LEMMA 5.3. *There exists a C^∞ change of basis of $\mathbb{C}^2$ such that, for all X close to X_0, $a_r(\widetilde{X})$ has purely imaginary coefficients.*

Proof. Let (f_1, f_2) be the canonical basis of $\mathbb{C}^2$. For X close to X_0, define

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B_r(X) = a_r(X) - a_r(X_0).$$

Since $B_r(X_0)$ is zero, the couple of vectors $(f_1' = (N - iB_r(\widetilde{X}))\, f_2, f_2)$ forms a basis of $\mathbb{C}^2$ for X close to X_0, and f_1' is a C^∞ vector-valued function of X. In this new basis, $N - iB_r(\widetilde{X})$ reads

$$\begin{pmatrix} b_1 & 1 \\ b_2 & 0 \end{pmatrix},$$

and the characteristic polynomial of $N - iB_r(\widetilde{X})$ is therefore

$$P(\xi) = \xi^2 - b_1\xi - b_2.$$

We also have the relation $N - iB_r(\widetilde{X}) = -ia_r(\widetilde{X}) - \kappa_r I_2$, from which we get

$$P(\xi) = \det\left[ia_r(\widetilde{X}) + (\kappa_r + \xi)I_2\right],$$

and relation (5.3) asserts that P has real coefficients. This completes the proof. $\square$

6. Concluding remarks. In both problems detailed in this paper, a weak stability result has been proved. Though the present study is just a constant coefficients analysis, it indicates the way to follow in order to get a nonlinear existence result. (We warn the reader that such a result is not guaranteed at the present time.)

Since both problems give rise to losses of derivatives on the solution of the corresponding linearized system, special attention should be paid when dealing with a variable coefficients linearized system. The usual linearized system (2.8) used in [25, 27, 30] is not appropriate in this case, since the right-hand side would involve some terms whose Sobolev norm needs to be controlled when one wants to construct an iterative scheme. Higher order terms in the Taylor expansion should therefore

be taken into account when linearizing equations (2.1) around a variable coefficients state $\overline{U}$.

It appears from Theorem 4.5 that the case of phase transitions in a van der Waals fluid is rather similar to the problem treated in [34]. The study of the variable linearized system should be carried out by using a parameter version of paradifferential calculus as introduced in [30].

To conclude, it is known since Majda's work that planar discontinuities for a multidimensional scalar conservation law are only weakly stable: since our method relies first on the elimination of the shock front, it cannot apply in the context of scalar conservation laws. Moreover, instability in this case is associated with the shock front symbol which is a second reason why we cannot deal with such equations. Nevertheless, we postpone the extension of the previous results to the general case of a multidimensional system to a future work.

Acknowledgments. The author is indebted to Sylvie Benzoni-Gavage and Guy Métivier for precious help and valuable comments.

REFERENCES

[1] M. S. Agranovič, *Boundary value problems for systems with a parameter*, Math. USSR-Sb., 13 (1971), pp. 25–64.
[2] V. I. Arnol'd, *Ordinary Differential Equations*, Springer-Verlag, Berlin, 1992.
[3] H. Beirão da Veiga, *On the existence theorem for the barotropic motion of a compressible inviscid fluid in the half-space*, Ann. Mat. Pura Appl. (4), 163 (1993), pp. 265–289.
[4] S. Benzoni-Gavage, *Stability of multi-dimensional phase transitions in a van der Waals fluid*, Nonlinear Anal., 31 (1998), pp. 243–263.
[5] S. Benzoni-Gavage, *Stability of subsonic planar phase boundaries in a van der Waals fluid*, Arch. Ration. Mech. Anal., 150 (1999), pp. 23–55.
[6] A. M. Blokhin, *Estimation of the energy integral of a mixed problem for gas dynamics equations with boundary conditions on the shock wave*, Sibirsk. Mat. Zh., 22 (1981), pp. 23–51, 229 (in Russian).
[7] A. M. Blokhin, *Uniqueness of the classical solution of a mixed problem for equations of gas dynamics with boundary conditions on a shock wave*, Sibirsk. Mat. Zh., 23 (1982), pp. 17–30, 222 (in Russian).
[8] J. Chazarain and A. Piriou, *Introduction to the Theory of Linear Partial Differential Equations*, North-Holland, Amsterdam, 1982.
[9] R. Courant and D. Hilbert, *Methods of Mathematical Physics, Vol. 2: Partial Differential Equations*, Interscience, New York, London, 1962.
[10] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Interscience, New York, 1948.
[11] W. Domański, *Surface and boundary waves for linear hyperbolic systems: Applications to basic equations of electrodynamics and mechanics of continuum*, J. Tech. Phys., 30 (1989), pp. 283–300.
[12] S. P. D'yakov, *On the stability of shock waves*, Ž. Èksper. Teoret. Fiz., 27 (1954), pp. 288–295 (in Russian).
[13] J. Erpenbeck, *Stability of step shocks*, Phys. Fluids, 5 (1962), pp. 1181–1187.
[14] H. Freistühler, *The persistence of ideal shock waves*, Appl. Math. Lett., 7 (1994), pp. 7–11.
[15] O. Guès, *Problème mixte hyperbolique quasi-linéaire caractéristique*, Comm. Partial Differential Equations, 15 (1990), pp. 595–645.
[16] R. Hersh, *Mixed problems in several variables*, J. Math. Mech., 12 (1963), pp. 317–334.
[17] R. L. Higdon, *Initial-boundary value problems for linear hyperbolic systems*, SIAM Rev., 28 (1986), pp. 177–217.
[18] D.-Y. Hsieh and X.-P. Wang, *Phase transition in van der Waals fluid*, SIAM J. Appl. Math., 57 (1997), pp. 871–892.
[19] V. M. Kontorovič, *Stability of shock waves in relativistic hydrodynamics*, Soviet Physics JETP, 34 (1958), pp. 127–132.
[20] H. O. Kreiss, *Initial boundary value problems for hyperbolic systems*, Comm. Pure Appl. Math., 23 (1970), pp. 277–298.

[21] P. D. Lax, *Hyperbolic systems of conservation laws.* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.
[22] P. D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 11, SIAM, Philadelphia, 1973.
[23] T. T. Li and W. C. Yu, *Boundary Value Problems for Quasilinear Hyperbolic Systems*, Duke University Mathematics Department, Durham, NC, 1985.
[24] A. Majda, *The existence of multi-dimensional shock fronts*, Mem. Amer. Math. Soc., 43 (1983).
[25] A. Majda, *The stability of multi-dimensional shock fronts*, Mem. Amer. Math. Soc., 41 (1983).
[26] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Springer-Verlag, New York, 1984.
[27] G. Métivier, *Problèmes mixtes non linéaires et stabilité des chocs multidimensionnels*, Astérisque, 152-153 (1987), pp. 37–53.
[28] G. Métivier, *The block structure condition for symmetric hyperbolic systems*, Bull. London Math. Soc., 32 (2000), pp. 689–702.
[29] S. Miyatake, *Neumann Operator for Wave Equation in a Half Space and Microlocal Orders of Singularities Along the Boundary*, Séminaire sur les Équations aux Dérivées Partielles, 1992–1993, Exp. No. XIV, École Polytechnique, Palaiseau, Cedex, France, 1993.
[30] A. Mokrane, *Problèmes mixtes hyperboliques non-linéaires*, Ph.D. Thesis, Université de Rennes I, Rennes Cedex, France, 1987.
[31] T. Ohkubo and T. Shirota, *On structures of certain L^2-well-posed mixed problems for hyperbolic systems of first order*, Hokkaido Math. J., 4 (1975), pp. 82–158.
[32] J. V. Ralston, *Note on a paper of Kreiss*, Comm. Pure Appl. Math., 24 (1971), pp. 759–762.
[33] R. D. Richtmyer, *Taylor instability in shock acceleration of compressible fluids*, Comm. Pure Appl. Math., 13 (1960), pp. 297–319.
[34] M. Sablé-Tougeron, *Existence pour un problème de l'élastodynamique Neumann non linéaire en dimension* 2, Arch. Ration. Mech. Anal., 101 (1988), pp. 261–292.
[35] P. Secchi, *Inflow-outflow problems for inviscid compressible fluids*, Commun. Appl. Anal., 2 (1998), pp. 81–110.
[36] D. Serre, *Systems of Conservation Laws.* 1, Cambridge University Press, Cambridge, UK, 1999.
[37] D. Serre, *Systems of Conservation Laws.* 2, Cambridge University Press, Cambridge, UK, 2000.
[38] M. Shearer, *Admissibility criteria for shock wave solutions of a system of conservation laws of mixed type*, Proc. Roy. Soc. Edinburgh Sect. A, 93 (1982/83), pp. 233–244.
[39] M. Slemrod, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Ration. Mech. Anal., 81 (1983), pp. 301–315.
[40] L. Truskinovsky, *About the "normal growth" approximation in the dynamical theory of phase transitions*, Contin. Mech. Thermodyn., 6 (1994), pp. 185–208.

SIAM J. MATH. ANAL.
Vol. 34, No. 1, pp. 173–203

SMOOTHNESS OF CENTER MANIFOLDS FOR MAPS AND FORMAL ADJOINTS FOR SEMILINEAR FDES IN GENERAL BANACH SPACES*

TERESA FARIA†, WENZHANG HUANG‡, AND JIANHONG WU§

Abstract. We develop a formal adjoint theory for retarded linear functional differential equations in Banach spaces and establish the existence and smoothness of center manifolds for nonlinearly perturbed equations. The hypotheses imposed here are significantly weaker than those that usually appear in the literature referring to semigroups for abstract functional differential equations, and the smoothness of the center manifolds for nonlinear perturbed equations is derived from our general results on the smoothness of center manifolds for maps in infinite-dimensional Banach spaces.

Key words. functional differential equations in Banach spaces, formal adjoint equations, resolvent compact operators, perturbations, center manifolds, smoothness

AMS subject classifications. 34K30, 34K06, 34K19, 34K17

PII. S0036141001384971

1. Introduction.

We consider the autonomous linear functional differential equations (FDEs) of retarded type,

$$\dot{u}(t) = A_T u(t) + L(u_t), \qquad u(t) \in X, \tag{1.1}$$

and the nonlinearly perturbed systems

$$\dot{u}(t) = A_T u(t) + L(u_t) + F(u_t), \tag{1.2}$$

where X is a Banach space, $r > 0$, $C := C([-r, 0]; X)$ is the Banach space of continuous mappings from $[-r, 0]$ to X with the sup norm, $u_t \in C$ is defined by $u_t(\theta) = u(t + \theta)$ for $t \in \mathbb{R}$ and $\theta \in [-r, 0]$, $L : C \longrightarrow X$ is a bounded linear operator, $A_T : D(A_T) \subset X \longrightarrow X$ is the infinitesimal generator of a compact C_0-semigroup of linear operators on X, and F is a sufficiently smooth nonlinear map with $F(0) = 0, DF(0) = 0$.

In the last two decades, there has been an increasing interest in retarded FDEs in Banach spaces. Typically, these equations depend on both spatial and temporal variables, with the time-dependence involving discrete or distributed delays. Such equations arise from a variety of situations in population dynamics and take the abstract form (1.1) or (1.2), where a diffusion term $d\Delta v(t, x)$ with $d = (d_1, \ldots, d_n) \in \mathbb{R}^n$ defines $A_T u(t) = d\Delta v(t, x)$ for $u(t)(x) := v(t, x), x \in \mathbb{R}^n$. See Wu [21] for more details.

*Received by the editors February 13, 2001; accepted for publication (in revised form) February 26, 2002; published electronically September 5, 2002.
http://www.siam.org/journals/sima/34-1/38497.html

†Departamento de Matemática, Faculdade de Ciências/CMAF, Universidade de Lisboa, 1749-016 Lisboa, Portugal (tfaria@lmc.fc.ul.pt). The work of this author was partially supported by FCT (Portugal) under project POCTI/32931/MAT/2000.

‡Department of Mathematical Sciences, University of Alabama in Huntsville, Huntsville, AL 35899 (huang@ultra.math.uah.edu).

§Department of Mathematics and Statistics, York University, Toronto, ON, M3J 1P3 Canada (wujh@mathstat.yorku.ca). The work of this author was partially supported by the Natural Sciences and Engineering Research Council of Canada and by Canada Research Chairs Program.

The purpose of the present work is to establish two necessary technical tools—a formal adjoint theory for equations of type (1.1) and the existence and smoothness of center manifolds for nonlinearly perturbed equation (1.2)—in order to develop a normal form theory on invariant manifolds of (1.2).

Several extensions of the formal adjoint and invariant manifold theory for FDEs in $\mathbb{R}^n$ (see Hale [8]) to infinite-dimensional Banach spaces have been developed in different frameworks. Related to our present work is the paper of Travis and Webb [18], where the authors initiated a formal adjoint theory for linear equations of the form (1.1); other related work includes Arino and Sanchez [1], Huang [9], Nakagiri [13], Schumacher [15], Shin and Naito [16], Wu [21], and Yamamoto and Nakagiri [22], to mention a few. We should particularly remark that a quite complete theory has also been developed for FDEs in Banach spaces of type (1.1) and (1.2) regarding duality, formal adjoint theory, and invariant manifolds (cf., e.g., Memory [12], Lin, So, and Wu [11], Wu [21], and Faria [5]) under some quite severe constraints. In fact, assume that the eigenvectors of A_T form a basis for X in the following sense: if $\mu_k, k \in \mathbb{N}$, are the eigenvalues of A_T with associated eigenvectors $\beta_k, k \in \mathbb{N}$, then every $x \in X$ is written in a unique way as $x = \sum_{k\in\mathbb{N}} x_k$, where $x_k \in span\{\beta_k\}, k \in \mathbb{N}$, with $A_T x = \sum_{k\in\mathbb{N}} \mu_k x_k$. Assume also that $L(\varphi\beta_k) \in span\{\beta_k\}$ for all $\varphi \in C([-r, 0]; \mathbb{R})$ and all eigenvectors β_k. Then it is possible to decompose the characteristic equation of the abstract FDE into a sequence of characteristic equations in $\mathbb{R}$. This decomposition yields a decomposition of (1.1) into a sequence of scalar FDEs, to which the standard formal adjoint theory for FDEs in $\mathbb{R}^n$ of Hale [8] can be applied (see [11], [12], [21], and other references therein). A slightly weaker hypothesis was considered in [5], as follows. In addition to the assumption that the eigenvectors of A_T form a basis for X, suppose now that the set of eigenvalues of A_T can be written as $\{\mu_k^{i_k} : k \in \mathbb{N}, i_k = 1, \dots, p_k\}$; for each $k \in \mathbb{N}$, let B_k be the generalized eigenspace for A_T associated with the block of eigenvalues $\{\mu_k^{i_k} : i_k = 1, \dots, p_k\}$, and assume that $L(\mathcal{B}_k) \subset B_k$, where $\mathcal{B}_k = \{\varphi \in C : \varphi(\theta) \in B_k \text{ for } \theta \in [-r, 0]\}$. This means that the eigenvalues of A_T can be organized by blocks in such a way that L does not mix the modes of the generalized eigenspaces associated with the eigenvalues in each block. Under these conditions, (1.1) is decomposed into a sequence of FDEs in finite-dimensional spaces (whose dimensions are now equal to the dimensions of the generalized eigenspaces B_k associated with each block $\{\mu_k^{i_k} : i_k = 1, \dots, p_k\}$), and again one can apply the adjoint theory for FDEs in $\mathbb{R}^n$. However, these hypotheses impose severe restrictions on the applicability of the approach to a wide range of problems arising from population dynamics. For instance, even if A_T is an n-dimensional elliptic operator with $n > 1$, it is unknown whether the eigenfunctions of A_T form a basis of X. Moreover, the above assumption that the linear operator L does not mix the modes of the eigenfunction spaces of the operator A_T is not realistic, for this almost implies that the operator L is a scalar multiplication.

Our goal is to develop a complete formal adjoint theory and center manifold theory without the aforementioned restrictions. The main sources of inspiration for our work on adjoint theory presented here are the work of Travis and Webb [18] for (1.1) and the work of Arino and Sanchez [1] for equations of the form $\dot{u}(t) = L(u_t)$, with $L : C \longrightarrow X$ being a bounded linear operator. More specifically, Travis and Webb [18] set the basis for an adjoint theory by introducing an adequate bilinear form $\langle\langle \cdot, \cdot \rangle\rangle$, which serves as the formal duality between C and its dual C^*, as well as an adequate definition of formal adjoint equation for (1.1). However, their theory was not completed in the following sense: in order to set a suitable framework to construct normal forms for

perturbed FDE (1.2), a formal adjoint theory should eventually provide an analytic formula for the decomposition of the phase space C by a nonempty finite set Λ of characteristic values for (1.1). Here, we present results that enable us to decompose C by Λ as the direct sum $C = P \oplus Q$, where P is the generalized eigenspace associated with Λ and $Q = \{\varphi \in C : \langle\langle \psi, \varphi \rangle\rangle = 0 \text{ for all } \psi \in P^*\}$, where P^* is the generalized eigenspace associated with Λ for the formal adjoint equation.

Since we deal with infinite-dimensional Banach spaces X, rather than finite-dimensional ones, our main difficulty is to use the formal duality to relate the generalized eigenspaces of the infinitesimal generator for the semigroup induced by the solutions of (1.1) with the generalized eigenspaces of its formal adjoint. Without having to impose further hypotheses on X or on the operators A_T and L, we succeeded in expressing the kernel and range for these generalized eigenspaces in terms of the kernel and range for some auxiliary operators. (This is a generalization of the operators introduced by Hale [8] for the case $X = \mathbb{R}^n$.) It turns out that these auxiliary operators are crucial for deriving the decomposition $C = P \oplus Q$ by a nonempty finite set Λ of characteristic eigenvalues because, as we shall prove, they have compact resolvents and closed ranges.

For the sake of exposition, we include some definitions and results from [18]. But we should emphasize that some results about duality in [18] were proven under stronger hypotheses than the ones assumed in this paper. Namely, in the present work the Banach space X is not required to be reflexive; also in [18, Propositions 4.14 and 4.15], some conditions on the characteristic operator were imposed in order to derive some results, such as that the point spectra for the infinitesimal generator of the semigroup defined by the mild solutions of (1.1) and for its formal adjoint coincide. Our techniques and results on formal adjoints are different from those in [1] for equations of type $\dot{u}(t) = L(u_t)$ (i.e., where A_T is absent). In [1] the authors considered only elements in Λ that are not in the essential spectrum, so that their auxiliary operators are Fredholm operators, while in the present paper we prove that the corresponding auxiliary operators have compact resolvents and closed ranges (two key points in establishing a Fredholm alternative result) from which the decomposition $C = P \oplus Q$ is deduced. Also, potential applications of the results in the present paper are much different from those of [1]. For instance, as we have already mentioned, (1.1) includes reaction-diffusion equations with delays as special cases.

As mentioned above, our second goal is to obtain the existence and smoothness of the center manifold. We notice that center manifolds are of particular interest in applications since the qualitative behavior of the solutions of a nonlinear equation in a neighborhood of an equilibrium can be described by the flow on these manifolds. See, for example, Carr [3]. See also Vanderbauwhede and van Gils [20], Vanderbauwhede and Iooss [19], and Diekmann et al. [4] for the theory of center manifolds for FDEs in $\mathbb{R}^n$. As already observed in the aforementioned papers, the phase space for FDE (1.2) is a Banach space which does not admit a smooth cut-off function, and thus it is a very challenging task to obtain the smoothness of center manifolds. Such a difficult issue was addressed for FDEs in $\mathbb{R}^n$ by Vanderbauwhede and van Gils [20], and the details are presented by Diekmann et al. [4]. In the recent work of Krisztin, Walther, and Wu [10], the existence and C^1-smoothness of various invariant manifolds for C^1-maps in general Banach spaces were established. Here we utilize some of the ideas in [10] and prove general C^k-smoothness for C^k-maps, with k being an arbitrary positive integer, and we apply this general smoothness result for maps to obtain the existence and C^k-smoothness of center manifolds for the semiflow generated by (1.2). Such a general

smoothness result is necessary for the normal form theory to be developed later, as the normal forms usually involve Taylor series expansions of various nonlinear maps involved in the center manifold reduction.

Although our final goal is to use formal adjoints and center manifolds as basic tools to develop a normal form theory for equations in the form (1.2), we note that the results presented here are important by themselves, and a decomposition of the phase space for linear equations and center manifolds for semilinear equations could be applied in different frameworks of qualitative theory for FDEs.

The paper is organized as follows. In section 2, some definitions and results are recalled, most of them from [18]. Sections 3 and 4 address a complete formal adjoint theory for FDEs (1.1): the auxiliary operators are introduced in section 3, and we derive some important properties of their spectra and resolvents; in section 4, a Fredholm alternative result is presented, and the phase space C is decomposed by a nonempty finite set Λ of characteristic eigenvalues of (1.1) by using the formal adjoint equation. Section 5 develops general results for the smoothness of center-stable and center-unstable manifolds for maps in Banach spaces, and section 6 applies these results to obtain the existence and regularity of center manifolds for perturbed FDE (1.2) at the zero equilibrium.

Because of space limitations, other important properties of the center manifold, such as the local invariance and attractivity, will be studied in a separate paper.

We now list notation that will be used throughout the paper. For a given Banach space X and for a linear operator A from its domain in X to X, we shall use $D(A)$, $R(A)$, and $N(A)$ to denote the domain, range, and kernel of A, respectively. The spectrum, point spectrum, and resolvent of A are considered as subsets of $\mathbb{C}$ and are denoted by $\sigma(A)$, $\sigma_P(A)$, and $\rho(A)$, respectively. If $\lambda \in \sigma_P(A)$, then $\mathcal{M}_\lambda(A)$ is the generalized eigenspace associated with λ.

2. Preliminary results and definitions. Consider

$$\dot{u}(t) = A_T u(t) + L(u_t), \qquad t \geq 0, \quad u(t) \in X, \tag{2.1}$$

where X is a Banach space over the field $\mathbb{C}$, $r > 0$, $C := C([-r, 0]; X)$ is the Banach space of continuous mappings from $[-r, 0]$ to X with the sup norm, $L : C \longrightarrow X$ is a bounded linear operator, and $A_T : D(A_T) \subset X \longrightarrow X$ is linear. As usual, $u_t \in C$ denotes the shifted restriction of u to $[t - r, t]$, i.e., $u_t(\theta) = u(t + \theta)$ for $-r \leq \theta \leq 0$. We require the following assumptions:

(H1) A_T generates a C_0-semigroup of linear operators $\{T(t)\}_{t\geq 0}$ on X, with $\|T(t)\| \leq Me^{\omega t}$ $(t \geq 0)$ for some $M \geq 1$, $\omega \in \mathbb{R}$.

(H2) $T(t)$ is a compact operator for each $t > 0$.

For $u \in C([-r, \infty); X)$, u is said to be a *mild solution* of (2.1) with initial condition $\varphi \in C$ if it satisfies

$$\begin{cases} u(t) = T(t)\varphi(0) + \int_0^t T(t - s)L(u_s)ds, & t \geq 0, \\ u_0 = \varphi. \end{cases} \tag{2.2}$$

(See, e.g., [23, p. 75] for the definition of integral used here.) It is known that the initial value problem (2.2) has a unique solution denoted by $u(\varphi)(t), t \in [-r, \infty)$. Moreover, for the operators $U(t), t \geq 0$, given by

$$U(t) : C \longrightarrow C, \qquad U(t)\varphi = u_t(\varphi), \tag{2.3}$$

from Propositions 2.4, 3.1, and 3.2 in Travis and Webb [18], we have the following proposition.

PROPOSITION 2.1. *Assume* (H1). *Then* $\{U(t)\}_{t\geq 0}$ *is a* C_0*-semigroup of bounded linear operators on* C. *Its infinitesimal generator* $A_U : C \longrightarrow C$ *is given by*

$$\begin{aligned} &A_U\varphi = \dot{\varphi}, \\ &D(A_U) = \{\varphi \in C : \dot{\varphi} \in C, \varphi(0) \in D(A_T), \dot{\varphi}(0) = A_T\varphi(0) + L(\varphi)\}. \end{aligned} \tag{2.4}$$

Moreover, if (H2) *holds, then* $U(t)$ *is a compact operator for each* $t > r$.

Since $\{U(t)\}_{t\geq 0}$ is eventually compact (i.e., there exists $t_0 > 0$ such that $U(t)$ is a compact operator for every $t > t_0$), from Greiner [7, p. 209] the next result follows.

PROPOSITION 2.2. *Assume* (H1), (H2) *and let* A_U *be defined by* (2.4). *Then we have the following:*

(i) $\sigma(A_U) = \sigma_P(A_U)$ *and every* $\lambda \in \sigma(A_U)$ *is a pole of finite order of the resolvent* $R(\lambda; A_U) = (\lambda I - A_U)^{-1}$;
(ii) *for each* $\lambda \in \sigma(A_U)$, *the generalized eigenspace* $\mathcal{M}_\lambda(A_U)$ *is finite-dimensional;*
(iii) *for each* $\alpha \in \mathbb{R}$, *the set* $\{\lambda \in \sigma(A_U) : \mathrm{Re}\lambda \geq \alpha\}$ *is finite.*

From the general theory of C_0-semigroups and compact operators, we also conclude the following.

PROPOSITION 2.3. *Assume* (H1), (H2) *and let* $\lambda \in \mathbb{C}$. *If* $\lambda \in \sigma(A_U)$, *then the ascent and descent of* $A_U - \lambda I$ *are both equal to* m, *where* m *is the order of* λ *as a pole of the resolvent* $R(\lambda; A_U)$. *Furthermore,*

$$C = N[(A_U - \lambda I)^m] \oplus R[(A_U - \lambda I)^m], \tag{2.5}$$

where $N[(A_U - \lambda I)^m] = \mathcal{M}_\lambda(A_U)$ *and* $R[(A_U - \lambda I)^m]$ *is a closed subspace of* C.

Proof. The first part follows directly from Theorem V.10.1 of Taylor and Lay [17, p. 330]. Now, let $k \in \mathbb{N}$, $t > r$. Since $U(t)$ is compact, $N[(U(t) - \mu I)^k]$ is finite-dimensional for $\mu \in \sigma(U(t))$. On the other hand, from the general theory of C_0-semigroups,

$$N[(U(t) - \mu I)^k] = \bigoplus_{\lambda \in S_\mu} N[(A_U - \lambda I)^k], \quad \text{where } S_\mu = \{\lambda \in \sigma(A_U) : e^{\lambda t} = \mu\}.$$

Thus, for m the ascent of λ, $N[(A_U - \lambda I)^m] = \mathcal{M}_\lambda(A_U)$ is finite-dimensional and Theorem IV.5.10 of Taylor and Lay [17, p. 217] implies that $R[(A_U - \lambda I)^m]$ is closed. ∎

For $\lambda \in \mathbb{C}$, we say that λ is a *characteristic value* for (2.1) if λ satisfies the *characteristic equation* given by

$$\Delta(\lambda)x = 0, \qquad x \in D(A_T) \setminus \{0\}, \tag{2.6}$$

where $\Delta(\lambda) : D(A_T) \subset X \longrightarrow X$ is defined by

$$\Delta(\lambda)x := A_T x + L(e^{\lambda \cdot} x) - \lambda x, \qquad x \in D(A_T), \tag{2.7}$$

and $e^{\lambda \cdot} x \in C$ is given by $(e^{\lambda \cdot} x)(\theta) = e^{\lambda\theta} x$ for $\theta \in [-r, 0]$ and $x \in X$. It is easy to see that $\lambda \in \sigma(A_U)$ if and only if λ is a characteristic value for (2.1), in which case

$$N(A_U - \lambda I) = \{e^{\lambda \cdot} x : x \in N(\Delta(\lambda))\}.$$

Note also that for $\psi \in C$, the equation $\psi = (A_U - \lambda I)\varphi$ has a solution $\varphi \in D(A_U)$ if and only if there is a $b \in D(A_T)$ satisfying the equation

$$\Delta(\lambda)b = \psi(0) - L\left(\int_0^\theta e^{\lambda(\theta - \xi)}\psi(\xi)d\xi\right). \tag{2.8}$$

In this case, the solution φ of $\psi = (A_U - \lambda I)\varphi$ is given by

$$\varphi(\theta) = e^{\lambda\theta}b + \int_0^\theta e^{\lambda(\theta-\xi)}\psi(\xi)d\xi, \qquad \theta \in [-r, 0]. \tag{2.9}$$

Here and throughout the remainder of this paper, for the sake of simplicity, we abuse notation and write explicitly the value of $\varphi \in C$ at an arbitrary given $\theta \in [-r, 0]$ in the evaluation of $L(\varphi)$. Namely, $L(\int_0^\theta e^{\lambda(\theta-\xi)}\psi(\xi)d\xi)$ should be understood as the value of L acting on the mapping $[-r, 0] \ni \theta \mapsto \int_0^\theta e^{\lambda(\theta-\xi)}\psi(\xi)d\xi \in X$.

We now assume that the linear operator L can be expressed in integral form by means of a function of bounded variation:

(H3) There is $\eta : [-r, 0] \longrightarrow \mathcal{L}(X, X)$ of bounded variation such that

$$L(\varphi) = \int_{-r}^0 d\eta(\theta)\varphi(\theta), \qquad \varphi \in C,$$

where $\mathcal{L}(X, X)$ denotes the Banach space of bounded linear operators from X into X.

Following Travis and Webb [18], we define the formal duality, the formal adjoint operator of L, and the formal adjoint equation of (2.1) below.

Let X^* be the dual of X and $C^* := C([0, r]; X^*)$. The *formal duality* between C^* and C is the bilinear form $\langle\langle\cdot, \cdot\rangle\rangle$ from $C^* \times C$ to the scalar field, defined by

$$\langle\langle\alpha, \varphi\rangle\rangle = \langle\alpha(0), \varphi(0)\rangle - \int_{-r}^0 \int_0^\theta \langle\alpha(\xi - \theta), d\eta(\theta)\varphi(\xi)\rangle d\xi \tag{2.10}$$

for $\alpha \in C^*, \varphi \in C$, where $\langle\cdot, \cdot\rangle$ is the usual duality between X^* and X. For $f \in C([0, r]; \mathbb{R})$ and $u^* \in X^*$, we use fu^* to denote $f \otimes u^*$ in C^*, i.e., $(fu^*)(s) = f(s)u^*$ for $0 \le s \le r$. We remark that

$$\langle\langle fu^*, \varphi\rangle\rangle = \langle u^*, f(0)\varphi(0)\rangle - \left\langle u^*, L\left(\int_0^\theta f(\xi - \theta)\varphi(\xi)d\xi\right)\right\rangle. \tag{2.11}$$

To avoid possible confusion, throughout this paper we adopt the following notation: given a densely defined linear operator B in a Banach space, we denote by B^* the (true) adjoint of B, also called the dual of B; and by ${}^*B$ we denote the formal adjoint of B relative to the formal duality $\langle\langle\cdot, \cdot\rangle\rangle$ defined above, in a sense that will soon be more clearly defined. The *formal adjoint operator* ${}^*L$ of L is given by

$${}^*L : C^* \longrightarrow X^*, \qquad {}^*L(\alpha) = \int_{-r}^0 d\eta^*(\theta)\alpha(-\theta), \tag{2.12}$$

where $\eta^*(\theta)$ is the adjoint of $\eta(\theta)$. Since η is of bounded variation, its adjoint operator $\eta^* : [-r, 0] \longrightarrow \mathcal{L}(X^*, X^*)$ is also of bounded variation. For (2.1), the *formal adjoint equation* is defined as

$$\dot\alpha(t) = -A_T^*\alpha(t) - {}^*L(\alpha^t), \qquad t \le 0, \tag{2.13}$$

where A_T^* is the adjoint of A_T and $\alpha^t \in C^*$ is given by $\alpha^t(s) = \alpha(t + s)$ for $s \in [0, r]$.

Consider the *mild solution* $\alpha^t(\psi)$ for (2.13) with initial condition $\psi \in C^*$, i.e., the solution of the integral equation

$$\begin{cases} \alpha(t) = T^*(-t)\psi(0) + \int_0^t T^*(s - t){}^*L(\alpha^s)ds, & t \le 0, \\ \alpha^0(\psi) = \psi. \end{cases}$$

As for (2.1), equation (2.13) generates a C_0-semigroup of linear operators $\{{}^*U(t)\}_{t\geq 0}$ on C^* defined by ${}^*U(t)\psi = \alpha^{-t}(\psi)$, whose infinitesimal generator ${}^*A_U$ is given by

$$\begin{aligned} &{}^*A_U\alpha = -\dot{\alpha}, \\ &D({}^*A_U) = \{\alpha \in C^* : \dot{\alpha} \in C^*, \alpha(0) \in D(A_T^*), -\dot{\alpha}(0) = A_T^*\alpha(0) + {}^*L(\alpha)\} \end{aligned} \tag{2.14}$$

and has the following properties (see Travis and Webb [18]):

$$\langle\langle {}^*A_U\alpha, \varphi\rangle\rangle = \langle\langle \alpha, A_U\varphi\rangle\rangle \quad \text{for } \alpha \in D({}^*A_U),\ \varphi \in D(A_U), \tag{2.15}$$

$$\langle\langle \alpha, \varphi\rangle\rangle = 0 \quad \text{for } \alpha \in N({}^*A_U - \mu I),\ \varphi \in N(A_U - \lambda I),\ \text{with } \lambda \neq \mu. \tag{2.16}$$

Note that (2.15) justifies the designation of ${}^*A_U$ as the formal adjoint of A_U, since its behavior relative to the formal duality $\langle\langle \cdot, \cdot\rangle\rangle$ is similar to the behavior of the (true) adjoint of an operator relative to the usual duality between a Banach space and its dual.

3. The point spectrum of ${}^*A_U$. The classic (formal) adjoint theory for FDEs in $\mathbb{R}^n$ will now be generalized to FDEs in Banach spaces, completing the theory initiated by Travis and Webb [18] and following the ideas of Arino and Sanchez [1], Busenberg and Huang [2], and Huang [9].

Similarly to what is done in section 7.3 of Hale [8] (see also [1]), we introduce some auxiliary operators that allow us to express the null space and range for $(A_U - \lambda I)^m,\ \lambda \in \mathbb{C}, m \in \mathbb{N}$, in terms of the null space and range of those auxiliary operators. For $\lambda \in \mathbb{C}, j \in \mathbb{N}_0, m \in \mathbb{N}$, we define the following linear operators:

$$L_\lambda^j : X \longrightarrow X, \quad L_\lambda^j(x) = L\left(\frac{\theta^j}{j!}e^{\lambda\theta}x\right), \tag{3.1}$$

$$\mathcal{L}_\lambda^{(m)} : [D(A_T)]^m \longrightarrow X^m, \quad \mathcal{L}_\lambda^{(m)} = \begin{pmatrix} \Delta(\lambda) & L_\lambda^1 - I & L_\lambda^2 & \dots & L_\lambda^{m-1} \\ 0 & \Delta(\lambda) & L_\lambda^1 - I & \dots & L_\lambda^{m-2} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \Delta(\lambda) & L_\lambda^1 - I \\ 0 & 0 & \dots & 0 & \Delta(\lambda) \end{pmatrix}, \tag{3.2}$$

$$\mathcal{R}_\lambda^{(m)} : C \longrightarrow X^m, \quad \mathcal{R}_\lambda^{(m)}(\psi) = \begin{pmatrix} -L\Big(\int_0^\theta e^{\lambda(\theta-\xi)}\frac{(\theta-\xi)^{m-1}}{(m-1)!}\psi(\xi)d\xi\Big) \\ \vdots \\ -L\Big(\int_0^\theta e^{\lambda(\theta-\xi)}(\theta-\xi)\psi(\xi)d\xi\Big) \\ \psi(0) - L\Big(\int_0^\theta e^{\lambda(\theta-\xi)}\psi(\xi)d\xi\Big) \end{pmatrix}. \tag{3.3}$$

With the definitions above, it is clear that $\Delta(\lambda) = \mathcal{L}_\lambda^{(1)} = A_T + L_\lambda^0 - \lambda I$. Moreover, from (2.8) and (2.9) it follows that $\psi \in R(A_U - \lambda I)$ if and only if there exists $b \in D(A_T)$ such that $\Delta(\lambda)b = \mathcal{R}_\lambda^{(1)}(\psi)$.

As in [1] and [8], we can carry out direct computations to obtain an explicit characterization of the spaces $N[(A_U - \lambda I)^m],\ R[(A_U - \lambda I)^m],\ m \in \mathbb{N}$. So we state the following proposition without a proof.

PROPOSITION 3.1. *Assume* (H1), (H2) *and let* $\lambda \in \mathbb{C}, m \in \mathbb{N}$. *Then*
(i) $\varphi \in N[(A_U - \lambda I)^m]$ *if and only if*

$$\varphi(\theta) = \sum_{j=0}^{m-1} \frac{\theta^j}{j!} e^{\lambda\theta} u_j, \qquad \theta \in [-r, 0], \qquad \text{with} \quad \begin{pmatrix} u_0 \\ \vdots \\ u_{m-1} \end{pmatrix} \in N(\mathcal{L}_\lambda^{(m)});$$

(ii) $\psi \in R[(A_U - \lambda I)^m]$ *if and only if* $\mathcal{R}_\lambda^{(m)}(\psi) \in R(\mathcal{L}_\lambda^{(m)})$.
From the definition of ${}^*L$ in (2.12), one can see that

$$\langle {}^*L(fu^*), u\rangle = \langle u^*, L(\hat{f}u)\rangle$$

for $u^* \in X^*, u \in X, f \in C([0, r]; \mathbb{R})$, where $\hat{f} \in C([-r, 0]; \mathbb{R})$ is given by $\hat{f}(\theta) := f(-\theta)$ for $\theta \in [-r, 0]$. Therefore, the adjoint $(L_\lambda^j)^*$ of L_λ^j ($j \in \mathbb{N}_0, \lambda \in \mathbb{C}$) is given by

$$(L_\lambda^j)^* u^* = {}^*L\left(\frac{(-\theta)^j}{j!} e^{-\lambda\theta} u^*\right), \qquad u^* \in X^*. \tag{3.4}$$

Similar to Proposition 3.1, we have an explicit characterization of $N[({}^*A_U - \lambda I)^m]$.
PROPOSITION 3.2. *Assume* (H1)–(H3). *For* $m \in \mathbb{N}, \lambda \in \mathbb{C}$,

$$\alpha \in N[({}^*A_U - \lambda I)^m] \quad \text{if and only if } \alpha(s) = \sum_{j=0}^{m-1} \frac{(-s)^j}{j!} e^{-\lambda s} x^*_{m-j-1}, \qquad s \in [0, r],$$

with $(x_0^*, \ldots, x_{m-1}^*)^T \in N((\mathcal{L}_\lambda^{(m)})^*)$. *In particular,* $\alpha \in N({}^*A_U - \lambda I)$ *if and only if* $\alpha(s) = e^{-\lambda s} x^*, s \in [0, r]$, *with* $x^* \in N(\Delta(\lambda)^*)$.

Proof. We have

$$(\mathcal{L}_\lambda^{(m)})^* = \begin{pmatrix} \Delta(\lambda)^* & 0 & \ldots & 0 \\ (L_\lambda^1)^* - I & \Delta(\lambda)^* & \ldots & 0 \\ (L_\lambda^2)^* & (L_\lambda^1)^* - I & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ (L_\lambda^{m-1})^* & \ldots & (L_\lambda^1)^* - I & \Delta(\lambda)^* \end{pmatrix},$$

with $(L_\lambda^j)^*$ given by (3.4). Using this and direct computations in the same spirit as in section 7.3 in Hale [8], we can complete the verification of Proposition 3.2. □

Now, we want to present a Fredholm alternative result relative to the formal adjoint. The following lemmas will establish some properties of the operators $\mathcal{L}_\lambda^{(m)}$ that will play an important role in this setting.

LEMMA 3.3. *Assume* (H1), (H2) *and let* $\lambda \in \mathbb{C}$. *Then* $\lambda \in \rho(A_U)$ *if and only if* $0 \in \rho(\Delta(\lambda))$.

Proof. For $\lambda \in \mathbb{C}$, it has been shown in section 2 that $\lambda \in \rho(A_U)$ if and only if $N(\Delta(\lambda)) = \{0\}$. On the other hand, $\Delta(\lambda) = A_T + L_\lambda^0 - \lambda I$, where A_T generates a compact C_0-semigroup of bounded linear operators and $L_\lambda^0 - \lambda I$ is linear and bounded. Hence, $\Delta(\lambda)$ is also the infinitesimal generator of a compact C_0-semigroup (see Proposition III.1.4 of Pazy [14, p. 79]). From the note in p. 51 of the same book, it follows that $0 \in \rho(\Delta(\lambda))$ if and only if 0 is not an eigenvalue of $\Delta(\lambda)$, or, equivalently, if and only if $N(\Delta(\lambda)) = \{0\}$. □

LEMMA 3.4. *Assume* (H1), (H2) *and let* $\lambda \in \mathbb{C}$ *and* $m \in \mathbb{N}$. *Then*

(i) *if* $\mu \in \rho(\Delta(\lambda))$, *then* $\mu \in \rho(\mathcal{L}_\lambda^{(m)})$ *and* $(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1}$ *is a compact operator;*

(ii) $R(\mathcal{L}_\lambda^{(m)})$ *is a closed subspace of* X^m.

Proof. The proof of (i) is given by induction. For $m = 1$, $\mathcal{L}_\lambda^{(1)} = \Delta(\lambda)$. We have already observed that $\Delta(\lambda)$ is the infinitesimal generator of a compact C_0-semigroup. Hence, for $\mu \in \rho(\Delta(\lambda))$ the resolvent $[\Delta(\lambda) - \mu I]^{-1}$ is compact (see Theorem II.3.3 of Pazy [14, p. 48]).

We now consider $\lambda \in \mathbb{C}, \mu \in \rho(\Delta(\lambda))$ and suppose that (i) is true for m. Since

$$\mathcal{L}_\lambda^{(m+1)} - \mu I = \begin{pmatrix} \mathcal{L}_\lambda^{(m)} - \mu I & \begin{pmatrix} L_\lambda^m \\ \vdots \\ L_\lambda^2 \\ L_\lambda^1 - I \end{pmatrix} \\ O & \Delta(\lambda) - \mu I \end{pmatrix},$$

$$(\mathcal{L}_\lambda^{(m+1)} - \mu I)^{-1} = \begin{pmatrix} (\mathcal{L}_\lambda^{(m)} - \mu I)^{-1} & -(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1} \begin{pmatrix} L_\lambda^m \\ \vdots \\ L_\lambda^2 \\ L_\lambda^1 - I \end{pmatrix} (\Delta(\lambda) - \mu I)^{-1} \\ O & (\Delta(\lambda) - \mu I)^{-1} \end{pmatrix}$$

exists and is bounded. Now, let $(y_n) \subset X^m, (z_n) \subset X$ be bounded sequences. The compactness of the operators $(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1}$ and $(\Delta(\lambda) - \mu I)^{-1}$ implies that there are subsequences $(y_{n_k}), (z_{n_k})$ such that

$$(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1} y_{n_k} \to w \in X^m, \qquad (\Delta(\lambda) - \mu I)^{-1} z_{n_k} \to x \in X.$$

Then $(\mathcal{L}_\lambda^{(m+1)} - \mu I)^{-1} \binom{y_{n_k}}{z_{n_k}}$ converges, proving that $(\mathcal{L}_\lambda^{(m+1)} - \mu I)^{-1}$ is a compact operator.

To prove (ii), let $(x_n) \subset [D(A_T)]^m$, $\mathcal{L}_\lambda^{(m)} x_n \to y \in X^m$. For $\mu \in \rho(\Delta(\lambda)), \mu \neq 0$,

$$\left[\frac{I}{\mu} + (\mathcal{L}_\lambda^{(m)} - \mu I)^{-1}\right] x_n = \frac{1}{\mu}(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1} \mathcal{L}_\lambda^{(m)} x_n \to \frac{1}{\mu}(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1} y.$$

The space $R[\frac{I}{\mu} + (\mathcal{L}_\lambda^{(m)} - \mu I)^{-1}]$ is closed, because $(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1}$ is compact (see Theorem V.7.8 of Taylor and Lay [17, p. 300]). Thus, there exists $x \in X^m$ such that

$$\frac{1}{\mu}(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1} y = \left[\frac{I}{\mu} + (\mathcal{L}_\lambda^{(m)} - \mu I)^{-1}\right] x,$$

i.e., $\mathcal{L}_\lambda^{(m)} x = y \in R(\mathcal{L}_\lambda^{(m)})$. □

The characterization of the point spectrum of ${}^*A_U$ relies on the next lemma.

LEMMA 3.5. *Assume* (H1)–(H3). *Consider* $\lambda \in \mathbb{C}, m \in \mathbb{N}$. *Then*

$$\dim N(\mathcal{L}_\lambda^{(m)}) = \dim N((\mathcal{L}_\lambda^{(m)})^*).$$

Proof. We may assume that $\lambda \in \sigma(A_U)$, i.e., $0 \in \sigma(\Delta(\lambda))$ (cf. Lemma 3.3). For $\mu \in \rho(\Delta(\lambda))$, then $\mu \in \rho(\mathcal{L}_\lambda^{(m)})$ by Lemma 3.4, and we conclude that

$$N(\mathcal{L}_\lambda^{(m)}) = N\left((\mathcal{L}_\lambda^{(m)} - \mu I)^{-1} + \frac{I}{\mu}\right),$$
$$N((\mathcal{L}_\lambda^{(m)})^*) = N\left([(\mathcal{L}_\lambda^{(m)})^* - \mu I]^{-1} + \frac{I}{\mu}\right).$$

Since $\mathcal{L}_\lambda^{(m)}$ is densely defined, we also conclude that $\mu \in \rho((\mathcal{L}_\lambda^{(m)})^*)$ and $[(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1}]^* = [(\mathcal{L}_\lambda^{(m)})^* - \mu I]^{-1}$ (cf. Lemma I.10.2 of Pazy [14, p. 38]). It remains to be proved that $N((\mathcal{L}_\lambda^{(m)} - \mu I)^{-1} + \frac{I}{\mu})$ and $N([(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1}]^* + \frac{I}{\mu})$ have the same dimension. Since $(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1}$ is a compact operator, so is its adjoint $[(\mathcal{L}_\lambda^{(m)} - \mu I)^{-1}]^*$, and the result now follows from Theorem V.7.14 of Taylor and Lay [17, p. 303]. □

As an immediate and most relevant consequence of this lemma, we can now derive the following result.

PROPOSITION 3.6. *Assume* (H1)–(H3). *Then*

(i) $\sigma_P(A_U) = \sigma_P({}^*A_U)$;

(ii) $\dim N[(A_U - \lambda I)^m] = \dim N[({}^*A_U - \lambda I)^m], m \in \mathbb{N}$;

(iii) *the ascent of* $A_U - \lambda I$ *and* ${}^*A_U - \lambda I$ *are equal.*

Proof. Propositions 3.1 and 3.2 and Lemma 3.5 imply (ii), from which (i) and (iii) follow. □

Remark 3.1. We note that (i) of Proposition 3.6 was proven in Proposition 4.14 of Travis and Webb [18] under the additional hypothesis $N(\Delta(\lambda)) \neq \{0\}$ if and only if $N(\Delta(\lambda)^*) \neq \{0\}$.

Remark 3.2. In the literature dealing with adjoint semigroups for FDEs in Banach spaces (cf., e.g., Nakagiri [13] and Travis and Webb [18, p. 412]), it is often assumed that the Banach space X is reflexive in order to have nice properties for adjoint semigroups. Here, we are able to develop the adjoint theory without imposing such a condition. Of course, if this condition holds, further properties for ${}^*A_U$ and ${}^*U(t)$ are obtained. For example, if the Banach space X is reflexive, then the adjoint A_T^* of A_T is the infinitesimal generator of the adjoint C_0-semigroup $\{T(t)^*\}_{t\geq 0}$ (cf. Pazy [14, p. 39]). For $t > 0$, $T(t)$ is a compact operator, and hence its adjoint $T(t)^*$ is also compact. Since (H1) and (H2) are fulfilled with $A_T, T(t)$ replaced by $A_T^*, T(t)^*$, respectively, the conclusions of Propositions 2.1, 2.2, and 2.3 hold for ${}^*A_U, {}^*U(t)$ $(t > 0)$ instead of $A_U, U(t)$ $(t > 0)$. In particular, $\sigma_P({}^*A_U) = \sigma({}^*A_U)$.

Remark 3.3. In Arino and Sanchez [1], a formal adjoint theory was established for equations of the form $\dot{u}(t) = L(u_t)$, where $L : C \longrightarrow X$ is a bounded linear operator. Since $A_T = 0$, the C_0-semigroup $\{U(t)\}_{t\geq 0}$ associated with the solutions of this equation is not eventually compact in general. For this reason, in [1] the authors restricted their study to eigenvalues of the infinitesimal generator that are not in the essential spectrum. With this restriction, the corresponding operators $\mathcal{L}_\lambda^{(m)}$ are Fredholm operators, instead of having compact resolvent. However, for our purposes and in view of applications, it is more interesting to consider equations of type (2.1) rather than $\dot{u}(t) = L(u_t)$, and in this situation no restrictions on the eigenvalues have to be assumed.

4. Decomposition of the phase space by using the formal adjoint theory. In this section, we always assume (H1)–(H3). The Fredholm alternative is stated

in the next result.

PROPOSITION 4.1. *Consider $\lambda \in \sigma(A_U)$ and $m \in \mathbb{N}$. Then $\psi \in R[(A_U - \lambda I)^m]$ if and only if $\langle\langle \alpha, \psi \rangle\rangle = 0$ for all $\alpha \in N[({}^*A_U - \lambda I)^m]$. In particular, $\psi \in R(A_U - \lambda I)$ if and only if*

$$\langle\langle e^{-\lambda \cdot} u^*, \psi \rangle\rangle = 0 \quad \text{for all } u^* \in N(\Delta(\lambda)^*).$$

Proof. Since $R(\mathcal{L}_\lambda^{(m)})$ is closed (Lemma 3.4), we have

$$R(\mathcal{L}_\lambda^{(m)}) = N((\mathcal{L}_\lambda^{(m)})^*)^\perp.$$

Thus, Proposition 3.1 implies that

$$\psi \in R[(A_U - \lambda I)^m] \quad \text{if and only if } \langle Y^*, \mathcal{R}_\lambda^{(m)}(\psi) \rangle = 0$$

for all $Y^* \in N((\mathcal{L}_\lambda^{(m)})^*)$. For $Y^* = (y_0^*, \dots, y_{m-1}^*)^T \in (X^*)^m$, from (2.11) and (3.3) we have

$$\begin{aligned}
&\langle Y^*, \mathcal{R}_\lambda^{(m)}(\psi) \rangle \\
&= -\sum_{j=0}^{m-1} \left\langle y_j^*, L\left(\int_0^\theta e^{\lambda(\theta-\xi)} \frac{(\theta-\xi)^{m-j-1}}{(m-j-1)!} \psi(\xi) d\xi \right) \right\rangle + \langle y_{m-1}^*, \psi(0) \rangle \\
&= \sum_{j=0}^{m-1} \left\langle \left\langle e^{-\lambda s} \frac{(-s)^{m-j-1}}{(m-j-1)!} y_j^*, \psi \right\rangle \right\rangle,
\end{aligned}$$

and the result follows from Proposition 3.2. □

We note that the above result was established in Proposition 4.15 of Travis and Webb [18] only for the particular situation $m = 1$ and with the additional hypothesis that $\Delta(\lambda)$ has a closed range. In Proposition 4.1, the most important case is the case m equal to the ascent of $A_U - \lambda I$. For $\lambda \in \sigma(A_U)$, denote by $\mathcal{M}_\lambda(A_U)$ and $\mathcal{M}_\lambda({}^*A_U)$ the generalized eigenspaces for A_U and ${}^*A_U$ associated with λ, respectively.

PROPOSITION 4.2. *Let $\lambda \in \sigma(A_U)$ and m be the ascent of $A_U - \lambda I$. Then $C = \mathcal{M}_\lambda(A_U) \oplus Q_\lambda$, with $\mathcal{M}_\lambda(A_U) = N[(A_U - \lambda I)^m]$, $\mathcal{M}_\lambda({}^*A_U) = N[({}^*A_U - \lambda I)^m]$, and*

$$Q_\lambda = \{\psi \in C : \langle\langle \alpha, \psi \rangle\rangle = 0 \quad \text{for all } \alpha \in \mathcal{M}_\lambda({}^*A_U)\}. \tag{4.1}$$

Proof. From Proposition 3.6, m is also the ascent of ${}^*A_U - \lambda I$. On the other hand, Proposition 4.1 implies that $\psi \in R[(A_U - \lambda I)^m]$ if and only if $\langle\langle \alpha, \psi \rangle\rangle = 0$ for all $\alpha \in \mathcal{M}_\lambda({}^*A_U)$. Decomposition (2.5) is therefore written as $C = \mathcal{M}_\lambda(A_U) \oplus Q_\lambda$, with $Q_\lambda = R[(A_U - \lambda I)^m]$ defined by (4.1). □

LEMMA 4.3. *For $\lambda, \mu \in \sigma(A_U)$, $\lambda \neq \mu$, and $m, r \in \mathbb{N}$,*

$$\langle\langle \alpha, \varphi \rangle\rangle = 0 \quad \text{for all } \alpha \in N[({}^*A_U - \lambda I)^m] \quad \text{and} \quad \varphi \in N[(A_U - \mu I)^r].$$

Proof. This lemma generalizes formula (2.16) for $m \in \mathbb{N}$. It relies on the identity (2.15) and is easily verified by using arguments as in Lemma 9 of Arino and Sanchez [1], so we omit the details here. □

Let $\lambda \in \sigma(A_U)$ and choose bases

$$\Phi_\lambda = (\varphi_1, \dots, \varphi_{p_\lambda}), \qquad \Psi_\lambda = (\psi_1, \dots, \psi_{p_\lambda})^T$$

of $\mathcal{M}_\lambda(A_U)$ and $\mathcal{M}_\lambda({}^*A_U)$, respectively, where $p_\lambda = \dim \mathcal{M}_\lambda(A_U) = \dim \mathcal{M}_\lambda({}^*A_U)$. Define a $p_\lambda \times p_\lambda$ matrix

$$\langle\langle \Psi_\lambda, \Phi_\lambda \rangle\rangle := [\langle\langle \psi_i, \varphi_j \rangle\rangle]_{i,j=1,\dots,p_\lambda}.$$

Suppose that $\langle\langle \Psi, \Phi \rangle\rangle c = 0$ for some constant vector $c = (c_1, \dots, c_{p_\lambda})^T$. Then, $\langle\langle \alpha, c_1\varphi_1 + \cdots + c_{p_\lambda}\varphi_{p_\lambda} \rangle\rangle = 0$ for all $\alpha \in \mathcal{M}_\lambda({}^*A_U)$, and Proposition 4.2 implies that $c_1\varphi_1 + \cdots + c_{p_\lambda}\varphi_{p_\lambda} \in Q_\lambda \cap \mathcal{M}_\lambda(A_U) = \{0\}$ for Q_λ as in (4.1). This shows that $\langle\langle \Psi_\lambda, \Phi_\lambda \rangle\rangle$ is nonsingular. Therefore, we can always choose bases $\Psi_\lambda, \Phi_\lambda$ such that

$$\langle\langle \Psi_\lambda, \Phi_\lambda \rangle\rangle = I_{p_\lambda}, \qquad p_\lambda = \dim \mathcal{M}_\lambda(A_U). \tag{4.2}$$

If the bases are normalized in such a way that (4.2) is fulfilled, then there is a $p_\lambda \times p_\lambda$ constant matrix B_λ, with $\sigma(B_\lambda) = \{\lambda\}$, that satisfies simultaneously

$$\dot{\Phi}_\lambda = \Phi_\lambda B_\lambda \quad \text{and} \quad -\dot{\Psi}_\lambda = B_\lambda \Psi_\lambda. \tag{4.3}$$

Furthermore,

$$U(t) = \Phi_\lambda e^{B_\lambda t}, \qquad t > 0. \tag{4.4}$$

We are now in the position to decompose C by a finite set of characteristic eigenvalues of (2.1), using the formal duality $\langle\langle \cdot, \cdot \rangle\rangle$. Consider a nonempty finite set $\Lambda = \{\lambda_1, \dots, \lambda_s\} \subset \sigma(A_U)$ and define $\Phi_\Lambda = (\Phi_{\lambda_1}, \dots, \Phi_{\lambda_s}), \Psi_\Lambda = (\Psi_{\lambda_1}, \dots, \Psi_{\lambda_s})^T$, where $\Phi_{\lambda_j}, \Psi_{\lambda_j}$ are bases of the generalized eigenspaces $\mathcal{M}_{\lambda_j}(A_U), \mathcal{M}_{\lambda_j}({}^*A_U)$, respectively, such that (4.2) holds ($j = 1, \dots, s$). From Lemma 4.3, it follows that $\langle\langle \Psi_\Lambda, \Phi_\Lambda \rangle\rangle = I_p$, where $p = p_{\lambda_1} + \cdots + p_{\lambda_s}$.

PROPOSITION 4.4. *Assume* (H1)–(H3), *let* $\Lambda = \{\lambda_1, \dots, \lambda_s\} \subset \sigma(A_U)$, *define*

$$P_\Lambda = \mathcal{M}_{\lambda_1}(A_U) \oplus \cdots \oplus \mathcal{M}_{\lambda_s}(A_U),$$
$$P_\Lambda^* = \mathcal{M}_{\lambda_1}({}^*A_U) \oplus \cdots \oplus \mathcal{M}_{\lambda_s}({}^*A_U),$$

and consider bases $\Phi_\Lambda, \Psi_\Lambda$ *for* P_Λ, P_Λ^* *such that* $\langle\langle \Psi_\Lambda, \Phi_\Lambda \rangle\rangle = I_p$, $p = \dim P_\Lambda$. *Then there exists a subspace* Q_Λ *of* C, *invariant under* A_U *and* $U(t)$, $t \geq 0$, *such that*

$$C = P_\Lambda \oplus Q_\Lambda \tag{4.5}$$

with

$$Q_\Lambda = \{\varphi \in C : \langle\langle \Psi_\Lambda, \varphi \rangle\rangle = 0\}, \tag{4.6}$$

where $\langle\langle \Psi_\Lambda, \varphi \rangle\rangle := (\langle\langle \Psi_{\lambda_1}, \varphi \rangle\rangle, \dots, \langle\langle \Psi_{\lambda_s}, \varphi \rangle\rangle)^T$. *Moreover,* $\varphi \in C$ *is written according to decomposition* (4.6) *as* $\varphi = \varphi_{P_\Lambda} + \varphi_{Q_\Lambda}$, *where* $\varphi_{P_\Lambda} = \Phi_\Lambda \langle\langle \Psi_\Lambda, \varphi \rangle\rangle$ *and* $\varphi_{Q_\Lambda} \in Q_\Lambda$.

5. Center manifolds for maps in general Banach spaces: Smoothness. We start with the following general results on smooth center-stable manifolds for maps.

THEOREM 5.1. *Let* $f : U \to E$ *be a* C^1*-map on an open subset* U *of a Banach space* E *over* $\mathbb{R}$, *with a fixed point* p. *Let* $L = Df(p)$ *and assume that* E *has the following decomposition:*

$$E = E_s \oplus E_c \oplus E_u,$$

where E_s is a closed subspace, E_c and E_u are finite-dimensional, $L(E_s) \subset E_s$, $L(E_c) \subset E_c$, and $L(E_u) \subset E_u$. We further assume that

$$\sigma_s = \sigma(L|_{E_s} : E_s \to E_s) \text{ is contained in a compact subset of } \{z \in \mathbb{C} : |z| < 1\}$$

and

$$\begin{aligned}\sigma_c =& \sigma(L|_{E_c} : E_c \to E_c) \subset S^1_{\mathbb{C}},\\ \sigma_u =& \sigma(L|_{E_u} : E_u \to E_u) \subset \{z \in \mathbb{C} : |z| > 1\}.\end{aligned}$$

Let $E_{sc} = E_s \oplus E_c$. Then

(i) *there exist open neighborhoods N_{sc} of 0 in E_{sc}, N_u of 0 in E_u, N of p in U, and a C^1-map $w : N_{sc} \to E_u$ with $w(0) = 0$, $Dw(0) = 0$, and $w(N_{sc}) \subset N_u$ so that the shifted graph $W = p + \{z + w(z) : z \in N_{sc}\}$ satisfies $f(W \cap N) \subset W$ and $\cap_{n=0}^{\infty} f^{-n}(p + N_{sc} + N_u) \subset W$;*

(ii) *if f is C^k-smooth for an integer $k \geq 2$, then so is w.*

Part (i) was proved in [10]. Our argument for the general smoothness in (ii), given below, will be based on the following general C^1-smoothness result for fixed points of contractions depending on a parameter developed in [10].

LEMMA 5.2. *Let Y, Λ be Banach spaces over $\mathbb{R}$ and let an open set $P \subset \Lambda$, a map $h : Y \times P \to Y$, and a constant $\kappa \in [0, 1)$ be given with $|h(y, p) - h(\tilde{y}, p)| \leq \kappa |y - \tilde{y}|$ for all $y, \tilde{y}$ in Y and all $p \in P$. Consider a convex subset $M \subset Y$ and a map $\Phi : P \to M$ so that for every $p \in P$, $\Phi(p)$ is the unique fixed point of $h(\cdot, p) : Y \to Y$. Suppose the following hold:*

(i) *the restriction $h_0 = h|_{M \times P}$ has a partial derivative $D_2 h_0 : M \times P \to L(\Lambda, Y)$ and the map $D_2 h_0$ is continuous;*

(ii) *there are a Banach space Y_1 over $\mathbb{R}$ and a continuous injective linear map $j : Y \to Y_1$ so that the map $k = j \circ h_0$ is continuously differentiable with respect to Y in the sense that there is a continuous map $A : M \times P \to L(Y, Y_1)$ so that for every $(y, p) \in M \times P$ and every $\epsilon^* > 0$, there exists $\tilde{\delta} > 0$ with $|k(\tilde{y}, p) - k(y, p) - A(y, p)(\tilde{y} - y)| \leq \epsilon^* |\tilde{y} - y|$ for all $\tilde{y} \in M$ with $|\tilde{y} - y| \leq \tilde{\delta}$;*

(iii) *there exist maps $h^{(1)} : M \times P \to L(Y, Y)$ and $h_1^{(1)} : M \times P \to L(Y_1, Y_1)$ such that*

$$A(y, p)\hat{y} = j h^{(1)}(y, p)\hat{y} = h_1^{(1)}(y, p) j\hat{y} \quad \text{on } M \times P \times Y$$

and

$$|h^{(1)}(y, p)| \leq \kappa,\ |h_1^{(1)}(y, p)| \leq \kappa \quad \text{on } M \times P;$$

(iv) *the map $(y, p) \in M \times P \to j \circ h^{(1)}(y, p) \in L(Y, Y_1)$ is continuous.*

Then the map $j \circ \Phi : P \to Y_1$ is C^1-smooth and

$$D(j \circ \Phi)(p) = h_1^{(1)}(\Phi(p), p) \circ D(j \circ \Phi)(p) + j \circ D_2 h_0(\Phi(p), p) \quad \text{for all } p \in P.$$

For a given positive integer k and for given Banach spaces $Y_1, \ldots, Y_k$ and Y, let $\mathcal{L}^{(k)}(Y_1 \times \cdots \times Y_k, Y)$ be the Banach space of all continuous k-linear maps from $Y_1 \times \cdots \times Y_k$ to Y, equipped with the operator norm. If $Y_i = Y_1$ for all $1 \leq i \leq k$, we write $\mathcal{L}^{(k)}(Y_1, Y)$ for $\mathcal{L}^{(k)}(Y_1 \times \cdots \times Y_k, Y)$. Also, we will denote the kth derivative of a given map by D^k if it exists.

We now briefly recall some results and associated notation in [10] as a preparation for the proof of Theorem 5.1. Set $b = \inf_{\lambda \in \sigma_u} |\lambda|, a = \sup_{\lambda \in \sigma_s} |\lambda|$ and fix $\epsilon > 0$ with

$a+\epsilon<1<1+\epsilon<(1+\epsilon)^k<b-\epsilon$. Let P_s, P_c, P_u denote the projections of E onto E_s along $E_c\oplus E_u$, onto E_c along $E_s\oplus E_u$, and onto E_u along $E_c\oplus E_s$, respectively. Whenever convenient, we shall use abbreviations like

$$x_s=P_sx,\qquad x_c=P_cx,\qquad x_u=P_ux,\qquad P_{sc}=P_s+P_c,\qquad x_{cu}=x_c+x_u.$$

There exists a norm $|\cdot|$ on E which is equivalent to the originally given one and satisfies

$$\begin{aligned}&|x|=|x_s|+|x_c|+|x_u|,\\&|LP_sx|\le(a+\epsilon)|P_sx|,\\&|LP_cx|\le(1+\epsilon)|P_cx|,\\&|LP_ux|\ge(b-\epsilon)|P_ux|\end{aligned}$$

for all $x\in E$.

Set $V=U-p$. Consider the transformed map $g^*: x\in V\to f(x+p)-p\in E$ with fixed point 0 and $Dg^*(0)=L$. Define $r^*: V\to E$ as the nonlinear part of g^* by $r^*(x)=g^*(x)-Lx$, and then extend r^* to a map $r: E\to E$ by $r(x)=0$ for all $x\in E\setminus V$. Finally, let $g=L+r$.

To construct small Lipschitz continuous modifications of g which are smooth on strips containing the center-unstable space E_{cu}, we fix a norm $|\cdot|_{cu}$ on E_{cu} which is C^∞-smooth on $E_{cu}\setminus\{0\}$. The norm $||\cdot||: x\in E\to\max\{|x_s|,|x_{cu}|_{cu}\}\in\mathbb{R}$ is equivalent to $|\cdot|$. For $\delta>0$, set $E(\delta)=\{x\in E: ||x||<\delta\}$. Choose a C^∞-function $\rho:\mathbb{R}\to\mathbb{R}$ with $\rho([0,\infty))\subset[0,1], \rho(t)=1$ for $0\le t\le 1$, $\rho(t)=0$ for $t\ge 2$. For every $\delta>0$, define $r_\delta: E\to E$ by

$$r_\delta(x)=\rho\left(\frac{|x_{cu}|_{cu}}{\delta}\right)\rho\left(\frac{|x_s|}{\delta}\right)r(x)$$

and set $g_\delta=L+r_\delta$.

Fix $\delta_0>0$ so that $\overline{E(3\delta_0)}\subset V$ and that $r|_{E(3\delta_0)}$ is C^k-smooth and all lth derivatives, $1\le l\le k$, of $r|_{E(3\delta_0)}$ are bounded. Observing that for every $\delta\in(0,\delta_0)$ the restriction $r_\delta|_{\{x\in E:|x_s|<\delta\}}$ is given by $\rho(\frac{|x_{cu}|_{cu}}{\delta})r(x)$, it follows that $r_\delta|_{\{x\in E:|x_s|<\delta\}}$ is C^k-smooth and that the restriction of r_δ to $\{x\in E; |x_s|\le\frac{\delta}{2}\}$ has all lth derivatives bounded, $1\le l\le k$.

It was shown in [10] that there exist $\delta_1\in(0,\delta_0)$ and a nondecreasing function $\lambda:[0,\delta_1]\to[0,1]$ with $\lim_{\delta\to 0^+}\lambda(\delta)=0=\lambda(0)$ so that for each $\delta\in(0,\delta_1]$ and for all x,y in E, $|r_\delta(x)|\le\delta\lambda(\delta)$ and $|r_\delta(x)-r_\delta(y)|\le\lambda(\delta)|x-y|$.

For $\eta>0$, let E_η denote the Banach space of all sequences $\chi=(x_n)_0^\infty\in E^{\mathbb{N}}$ with

$$\sup_{j\in\mathbb{N}}|x_j|\eta^{-j}<\infty$$

and norm

$$||\chi||_\eta=\sup_{j\in\mathbb{N}}|x_j|\eta^{-j}.$$

Consider

$$\begin{cases}x_{n+1}=Lx_n+f_n & \text{for } n\ge 0,\\ P_{sc}x_0=z\end{cases}\tag{5.1}$$

for given $z\in E_{sc}$, $\phi=(f_n)_0^\infty\in E_\eta$, and $\eta\in(1+\epsilon,b-\epsilon)$.

Let $L_{sc} = L|_{E_{sc}} : E_{sc} \to E_{sc}$. It was shown in [10] that for fixed $z \in E_{sc}, 1+\epsilon < \eta < b-\epsilon$, and $\phi \in E_\eta$, if $\chi \in E_\eta$ satisfies (5.1), then

$$x_n = \sum_{j=0}^{n-1} L_{sc}^{n-j-1} P_{sc} f_j - \sum_{j=n}^{\infty} L_u^{n-j-1} P_u f_j + L_{sc}^n z \quad \text{for } n \geq 1$$

and

$$x_0 = z - \sum_{j=0}^{\infty} L_u^{-j-1} P_u f_j.$$

In particular, given $z \in E_{sc}$ and $\phi = (f_j)_0^\infty \in E_\eta$, there is at most one solution of (5.1) in E_η. Let

$$K : \{\chi \in E^{\mathbb{N}} : \chi \in E_\eta \text{ for some } \eta \in (1+\epsilon, \mathrm{b}-\epsilon)\} \to \mathrm{E}^{\mathbb{N}}$$

be given by

$$(K\phi)_n = \sum_{j=0}^{n-1} L_{sc}^{n-j-1} P_{sc} f_j - \sum_{j=n}^{\infty} L_u^{n-j-1} P_u f_j \quad \text{for } n \geq 1$$

and

$$(K\phi)_0 = -\sum_{j=0}^{\infty} L_u^{-j-1} P_u f_j.$$

Also, let

$$c(\eta) = \frac{1}{\eta - 1 - \epsilon} + \frac{1}{b - \epsilon - \eta}.$$

Then the linear map $K_\eta : E_\eta \to E_\eta$ given by $K_\eta \phi = K\phi$ is continuous with $|K_\eta| \leq c(\eta)$. Furthermore, for every $\eta \in (1+\epsilon, b-\epsilon)$, $z \in E_{sc}$, and $\phi \in E_\eta$, the sequence $\chi = K_\eta \phi + (L_{sc}^n z)_0^\infty \in E_\eta$ solves (5.1).

Consider the substitution operator

$$R_\delta : E^{\mathbb{N}} \to E^{\mathbb{N}} \quad \text{by } R_\delta(\chi) = (r_\delta(x_n))_0^\infty \quad \text{for } \chi = (x_n)_0^\infty \in E^{\mathbb{N}}.$$

For every $\eta \in (1+\epsilon, b-\epsilon)$, choose $\delta_\eta \in (0, \delta_1]$ with $\lambda(\delta_\eta) c(\eta) < 1$. Let $\eta \in (1+\epsilon, b-\epsilon)$ and $\delta \in (0, \delta_\eta)$. It was shown in [10] that $R_\delta(E_\eta) \subset E_\eta$, and the induced map $\gamma_{\delta\eta} : E_\eta \ni \chi \mapsto R_\delta(\chi) \in E_\eta$ is Lipschitz continuous with a Lipschitz constant $\lambda(\delta)$.

Therefore, for every $z \in E_{sc}$ and $\chi = (x_n)_0^\infty \in E_\eta$ the properties

$$x_{n+1} = g_\delta(x_n) \quad \text{for all } n \geq 0, \qquad P_{sc} x_0 = z$$

are equivalent to the fixed point equation $\chi = T_{\delta\eta}(\chi, z)$, where the map $T_{\delta\eta} : E_\eta \times E_{sc} \to E_\eta$ is given by

$$T_{\delta\eta}(\chi, z) = K_\eta(\gamma_{\delta\eta}(\chi)) + (L_{sc}^j z)_0^\infty.$$

As

$$|T_{\delta\eta}(\chi, z) - T_{\delta\eta}(\chi^*, z)|_\eta \leq c(\eta)\lambda(\delta)|\chi - \chi^*|_\eta$$

for all $\chi, \chi^* \in E_\eta$ and for all $z \in E_{sc}$, there is exactly one fixed point $\chi_{\delta\eta}(z) \in E_\eta$ of the contraction $T_{\delta\eta}(\cdot, z) : E_\eta \to E_\eta$ for every $z \in E_{sc}$. Moreover, $P_{sc}(\chi_{\delta\eta}(z))_0 = z$. In summary, $\chi \in E_\eta$ is a trajectory of g_δ with $P_{sc}x_0 = z$ if and only if $\chi = \chi_{\delta\eta}(z)$.

It was shown in [10] that the map $\chi_{\delta\eta} : z \in E_{sc} \to \chi_{\delta\eta}(z) \in E_\eta$ is Lipschitz continuous, and thus $w_{\delta\eta} : z \in E_{sc} \to P_u(\chi_{\delta\eta}(z))_0 \in E_u$ is Lipschitz continuous. To obtain the differentiability of $w_{\delta\eta}$, [10] proved the following important properties: if $0 < \delta < \delta_\eta$ and $\lambda(\delta) < \frac{(1-a-\epsilon)^2}{2}$, then for every $z \in E_{sc}$ with $|P_s z| < \frac{\delta}{2}$ and for all integers $j \geq 0$,

$$|P_s(\chi_{\delta\eta}(z))_j| < \frac{\delta}{2}. \tag{5.2}$$

We can now give the following proof.

Proof of Theorem 5.1. We divide the long proof into several steps. The first step concerns the proof of the C^1-smoothness. Except for the last remark, all results in Step 1 belong to [10].

Step 1. Fix $\eta, \tilde{\eta}, \overline{\eta}$ so that $1 + \epsilon < \eta < \tilde{\eta} \leq \overline{\eta}$ with $\overline{\eta} \in (\eta^k, b - \epsilon)$, and fix $\delta > 0$ so that

$$\delta < \delta_\eta, \qquad \lambda(\delta) < \frac{(1-a-\epsilon)^2}{2}, \qquad \kappa := \sup_{\tilde{\eta} \in [\eta, \overline{\eta}]} \lambda(\delta) c(\tilde{\eta}) < 1.$$

Let

$$P = \left\{ x \in E_{sc} : |x_s| < \frac{\delta}{2} \right\}.$$

P is an open set in the Banach space $\Lambda = E_{sc}$.

Recall that $r_\delta|_{\{x \in E : |x_s| < \delta\}}$ is C^k-smooth and $\sup\{|D^1 r_\delta(x)| : |x_s| < \delta\} \leq \lambda(\delta)$. It was shown in [10] that for any $\tilde{\eta} \in (\eta, \overline{\eta}]$, the linear map

$$A^{(1)}_{r_\delta}(\chi) : E^{\mathbb{N}} \ni \hat{\chi} = (\hat{x}_j)_0^\infty \mapsto (D^1 r_\delta(x_j)\hat{x}_j)_0^\infty \in E^{\mathbb{N}}, \quad \chi = (x_j)_0^\infty, \quad |P_s x_j| < \frac{\delta}{2}, \quad j \in \mathbb{N},$$

induces a continuous map $A^{(1)}_{r_\delta \tilde{\eta}\eta}$ from the convex set

$$M = \left\{ \chi \in E_\eta : |P_s x_j| < \frac{\delta}{2} \text{ for all } j \in \mathbb{N} \right\} \subset E_\eta$$

into $\mathcal{L}(E_\eta, E_{\tilde{\eta}})$.

Let $Y = E_\eta, h = T_{\delta\eta}|_{Y \times P}$. It is important to keep in mind that $\chi_{\delta\eta}(P) \subset M$. Define $\Phi : P \to M$ by $\Phi(z) = \chi_{\delta\eta}(z)$; we have $h(\Phi(p), p) = \Phi(p)$ for all $p \in P$. The map $h_0 = h|_{M \times P}$ is given by

$$h_0(\chi, z) = T_{\delta\eta}(\chi, z) = K(R_\delta(\chi)) + (L^j_{sc} z)_0^\infty,$$

so for every $(\chi, z) \in M \times P$ the derivative $D_2 h_0(\chi, z)$ exists and is given by

$$D_2 h_0(\chi, z)\tilde{z} = (L^j_{sc}\tilde{z})_0^\infty \in E_\eta.$$

This derivative is constant on $M \times P$ and therefore is continuous.

Set $Y_1 = E_{\tilde{\eta}}$ and define $j_{\tilde{\eta}\eta} : Y \to Y_1$ by

$$j_{\tilde{\eta}\eta}(\chi) = \chi.$$

Then $j_{\tilde{\eta}\eta}$ is continuous and injective, and the map $k = j_{\tilde{\eta}\eta} \circ h_0$ is given by

$$k(\chi, z) = T_{\delta_{\tilde{\eta}}}(\chi, z) = K_{\tilde{\eta}}(\gamma_{\delta_{\tilde{\eta}}}(\chi)) + (L^j_{sc} z)^\infty_0.$$

It was shown in [10] that the map $A : M \times P \ni (\chi, z) \mapsto K_{\tilde{\eta}} \circ A_{r_\delta \tilde{\eta}\eta}(\chi) \in L(Y, Y_1)$ is continuous, and each $A^{(1)}_{r_\delta}(\chi), \chi \in M$, defines elements

$$A^{(1)}_{r_\delta \eta\eta}(\chi) \in \mathcal{L}(Y, Y) \quad \text{with } |A^{(1)}_{r_\delta \eta\eta}(\chi)| \le \lambda(\delta)$$

and

$$A^{(1)}_{r_\delta \tilde{\eta}\tilde{\eta}}(\chi) \in \mathcal{L}(Y_1, Y_1) \quad \text{with } |A^{(1)}_{r_\delta \tilde{\eta}\tilde{\eta}}(\chi)| \le \lambda(\delta).$$

Define

$$h^{(1)} : M \times P \to \mathcal{L}(Y, Y) \quad \text{by } h^{(1)}(\chi, z) = K_\eta \circ A^{(1)}_{r_\delta \eta\eta}(\chi)$$

and

$$h^{(1)}_1 : M \times P \to \mathcal{L}(Y_1, Y_1) \quad \text{by } h^{(1)}_1(\chi, z) = K_{\tilde{\eta}} \circ A^{(1)}_{r_\delta \tilde{\eta}\tilde{\eta}}(\chi).$$

It was shown in [10] that

$$\max\{|h^{(1)}(\chi, z)|, |h^{(1)}_1(\chi, z)|\} \le \max\{c(\eta), c(\tilde{\eta})\}\lambda(\delta) = \kappa,$$

and all other conditions in Lemma 5.2 are satisfied. Therefore, $j_{\tilde{\eta}\eta} \circ \Phi = j_{\tilde{\eta}\eta} \circ (\chi_{\delta\eta}|_P)$ is C^1-smooth and $j_{\tilde{\eta}\eta} \circ \Phi = \chi_{\delta_{\tilde{\eta}}}|_P$. Moreover, $D^1(j_{\tilde{\eta}\eta} \circ \Phi)$ satisfies

$$D^1(j_{\tilde{\eta}\eta} \circ \Phi)(z) = K_{\tilde{\eta}} \circ A^{(1)}_{r_\delta \tilde{\eta}\tilde{\eta}}(\Phi(z)) \circ D^1(j_{\tilde{\eta}\eta} \circ \Phi)(z) + j_{\tilde{\eta}\eta} \circ (L^j_{sc} \cdot)^\infty_0, \qquad z \in P.$$

The final remark of this step is essential for the general smoothness to be proved in later steps. Recall that for any $\tilde{\eta} \in [\eta, \bar{\eta}]$, $K_{\tilde{\eta}} \circ A^{(1)}_{r_\delta \tilde{\eta}\tilde{\eta}}(\Phi(z)) \in \mathcal{L}(E_{\tilde{\eta}}, E_{\tilde{\eta}})$ and

$$|K_{\tilde{\eta}} \circ A^{(1)}_{r_\delta \tilde{\eta}\tilde{\eta}}(\Phi(z))|_{\mathcal{L}(E_{\tilde{\eta}}, E_{\tilde{\eta}})} \le c(\tilde{\eta})\lambda(\delta) \le \kappa < 1.$$

Therefore, $K_{\tilde{\eta}} \circ A^{(1)}_{r_\delta \tilde{\eta}\tilde{\eta}}(\Phi(z)) \in \mathcal{L}(E_{\tilde{\eta}}, E_{\tilde{\eta}})$ is a uniform contraction and the map

$$K_{\tilde{\eta}} \circ A^{(1)}_{r_\delta \tilde{\eta}\tilde{\eta}}(\Phi(z))L + j_{\tilde{\eta}\eta} \circ (L^j_{sc} \cdot)^\infty_0, \qquad z \in P,\ L \in \mathcal{L}(\Lambda, E_{\tilde{\eta}}),$$

has a unique fixed point $\Psi^{(1)}_{\tilde{\eta}}(z)$ in $\mathcal{L}(\Lambda, E_{\tilde{\eta}})$. Since $j_{\tilde{\eta}\eta} \circ \Psi^{(1)}_\eta(z) \in \mathcal{L}(\Lambda, E_{\tilde{\eta}})$, the uniqueness of a fixed point in $\mathcal{L}(\Lambda, E_{\tilde{\eta}})$ implies

$$\Psi^{(1)}_{\tilde{\eta}}(z) = j_{\tilde{\eta}\eta} \circ \Psi^{(1)}_\eta(z).$$

In particular,

$$D^1(j_{\tilde{\eta}\eta} \circ \Phi)(z) = \Psi^{(1)}_{\tilde{\eta}}(z) = j_{\tilde{\eta}\eta} \circ \Psi^{(1)}_\eta(z), \qquad z \in P.$$

Step 2. We now assume $k \geq 2$. For any given integer l with $1 \leq l \leq k$, consider the operator $A^{(l)}_{r_\delta}$ given by

$$A^{(l)}_{r_\delta}(\chi)(\chi^1, \ldots, \chi^l) = (D^l r_\delta(x_j)(x^1_j, \ldots, x^l_j))^\infty_0,$$
$$\chi = (x_j)^\infty_0, \quad \chi^i = (x^i_j)^\infty_0 \in E^{\mathbb{N}}, \qquad 1 \leq i \leq l.$$

Note that $A^{(l)}_{r_\delta}$ with $l = 1$ was introduced in Step 1. The operators $A^{(l)}_{r_\delta}$ with $1 \leq l \leq k$ are the substitution operators of $D^l r_\delta$; they can be regarded as the Nemytskii operators induced by $D^l r_\delta$ in the appropriate spaces.

As $r_\delta|_{\{z \in E; |z_s| \leq \frac{\delta}{2}\}}$ has all lth derivatives bounded, $1 \leq l \leq k$, we can show that

$$A^{(l)}_{r_\delta}(\chi)(E_{\eta^{r_1}} \times \cdots \times E_{\eta^{r_l}}) \subset E_{\eta^{r_1+\cdots+r_l}}, \qquad \chi \in M,\ 1 \leq r_i \leq l.$$

We are going to use induction on p with $1 \leq p \leq k$. (Note that for the remainder of this proof, p is not the fixed point of f.) The strategy is to show that the order of the smoothness of $j_{\tilde\eta\eta} \circ \Phi : P \to E_{\tilde\eta}$ is increased by at least one as $\tilde\eta$ passes η^{p-1}, from $(\eta, \eta^{p-1}]$ to $(\eta^{p-1}, \eta^p]$, and to construct higher order derivatives inductively.

Suppose $1 \leq p < k$ and suppose that for all integers q with $1 \leq q \leq p$ and for all $\tilde\eta \in [\eta^q, \bar\eta]$, the mapping $j_{\tilde\eta\eta} \circ \Phi : P \to E_{\tilde\eta}$ is C^q-smooth with

(i) $D^q(j_{\tilde\eta\eta} \circ \Phi) = j_{\tilde\eta\eta} \circ \Psi^{(q)}_\eta$;

(ii) $\Psi^{(q)}_\eta(z) \in \mathcal{L}^{(q)}(\Lambda, E_{\eta^q})$ as the unique solution of

$$F = KA^{(1)}_{r_\delta}(\Phi(z))F + H_q(z), \qquad F \in \mathcal{L}^{(q)}(\Lambda, E_{\eta^q}), \qquad z \in P,$$

with $H_1(z)\tilde z = (L^j_{sc}\tilde z)^\infty_0, \tilde z \in \Lambda$, and for $q \geq 2$,

$$H_q(z) = \sum_{2 \leq l \leq q, 1 \leq i \leq l, 1 \leq r_i \leq l, r_1+\cdots+r_l = q} KA^{(l)}_{r_\delta}(\Phi(z))(\Psi^{(r_1)}_\eta(z), \ldots, \Psi^{(r_l)}_\eta(z));$$

(iii) $j_{\tilde\eta\eta} \circ \Psi^{(q)}_\eta : P \to \mathcal{L}^{(q)}(\Lambda, E_{\tilde\eta})$ being continuous.

We want to show that the above statement is true for $q = p + 1$.

Step 3. Fix $\tilde\eta \in [\eta^{p+1}, \bar\eta]$ and let $X = \mathcal{L}^{(p)}(\Lambda, E_{\tilde\eta})$. For $F \in \mathcal{L}^{(p)}(\Lambda, E_{\eta^p})$ and $z \in P$, let

$$H(F, z) = KA^{(1)}_{r_\delta}(\Phi(z))F + H_p(z).$$

By the induction hypotheses in Step 2 and the estimates in Step 1, for any $\eta^* \in [\eta^p, \overline{\eta}]$, $F \in \mathcal{L}^{(p)}(\Lambda, E_{\eta^*}), z \in P$, we have $H(F, z) \in E_{\eta^*}$ and

$$|H(\tilde F, z) - H(F, z)| \leq c(\eta^*)\lambda(\delta)|\tilde F - F| \leq \kappa|\tilde F - F|, \qquad \tilde F, F \in \mathcal{L}^{(p)}(\Lambda, E_{\eta^*}).$$

Therefore, $H(\cdot, z)$ has a unique fixed point in $\mathcal{L}^{(p)}(\Lambda, E_{\eta^*})$. Note also that for $\eta^* = \eta^p$ this fixed point is given by $\Psi^{(p)}_\eta(z)$. From now on, we restrict $H : X \times P \to X$ and let $N = \mathcal{L}^{(p)}(\Lambda, E_{\eta^p})$, $H_0 = H|_{N \times P}$.

Step 4. Let $e_j : E^{\mathbb{N}} \to E$ be given by

$$e_j((z_i)^\infty_0) = z_j, \qquad (z_i)^\infty_0 \in E^{\mathbb{N}}.$$

Define $\Phi_j = e_j \circ \Phi : P \to E$ and $\Psi^{(l)}_{\eta j}(z)\tilde z = e_j \circ \Psi^{(l)}_\eta(z)\tilde z$ for $1 \leq l \leq p, z \in P$, and $\tilde z \in \Lambda$. We claim that Φ_j is C^1-smooth and $D\Phi_j(z)\tilde z = \Psi^{(1)}_{\eta j}(z)\tilde z$. In fact,

$\Phi_j = e_j \circ \Phi = e_j \circ j_{\tilde{\eta}\eta}\Phi$, and thus Φ_j is C^1-smooth since $j_{\tilde{\eta}\eta} \circ \Phi$ is. Moreover, $D(j_{\tilde{\eta}\eta} \circ \Phi) = j_{\tilde{\eta}\eta} \circ \Psi^{(1)}_\eta$, and thus

$$e_j(j_{\tilde{\eta}\eta}\Psi^{(1)}_\eta(z)\tilde{z}) = e_j D(j_{\tilde{\eta}\eta} \circ \Phi)(z)\tilde{z}.$$

This shows that $\Psi^{(1)}_{\eta j}(z)\tilde{z} = D\Phi_j(z)\tilde{z}$.

Step 5. We now prove that for any fixed $F \in \mathcal{L}^{(p)}(\Lambda, E_{\eta^p})$ and $\tilde{\eta} > \eta^{p+1}$, the mapping $P \ni z \mapsto KA^{(1)}_{r_\delta}(\Phi(z))F \in \mathcal{L}^{(p)}(\Lambda, E_{\tilde{\eta}})$ has a derivative, which is given by $KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)\cdot, F)$, and the map

$$P \times \mathcal{L}^{(p)}(\Lambda, E_{\eta^p}) \ni (z, F) \mapsto KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)\cdot, F) \in \mathcal{L}(\Lambda, \mathcal{L}^{(p)}(\Lambda, E_{\tilde{\eta}}))$$

is continuous.

Let

$$|D^l r_\delta|_\infty = \sup\left\{|D^l r_\delta(z)|; z \in E, |z_s| \le \frac{\delta}{2}\right\}.$$

Note that for $1 \le l \le k$, $|D^l r_\delta|_\infty < \infty$.

For any $z_i \in \Lambda$ with $1 \le i \le p$, let

$$F_j(z_1, \dots, z_p) = e_j(F(z_1, \dots, z_p)).$$

Then for $\tilde{z}, z \in P$ we have

$$\begin{aligned}
&\tilde{\eta}^{-j}|D^1 r_\delta(\Phi_j(\tilde{z}))F_j(z_1, \dots, z_p) - D^1 r_\delta(\Phi_j(z))F_j(z_1, \dots, z_p) \\
&\quad - D^2 r_\delta(\Phi_j(z))(\Psi^{(1)}_{\eta j}(z)(\tilde{z} - z), F_j(z_1, \dots, z_p))| \\
&\le \tilde{\eta}^{-j}|D^1 r_\delta(\Phi_j(\tilde{z})) - D^1 r_\delta(\Phi_j(z)) - D^2 r_\delta(\Phi_j(z))\Psi^{(1)}_{\eta j}(z)(\tilde{z} - z)|\eta^{pj}|F||z_1| \cdots |z_p|.
\end{aligned}$$

Therefore, for any $\epsilon > 0$ there exists an integer $J_0 \ge 0$ so that if $j \ge J_0$ and if $|\tilde{z} - z| \le 1$, then

$$\begin{aligned}
&\tilde{\eta}^{-j}|D^1 r_\delta(\Phi_j(\tilde{z}))F_j(z_1, \dots, z_p) - D^1 r_\delta(\Phi_j(z))F_j(z_1, \dots, z_p) \\
&\quad - D^2 r_\delta(\Phi_j(z))(\Psi^{(1)}_{\eta j}(z)(\tilde{z} - z), F_j(z_1, \dots, z_p))| \\
&\le [(\tilde{\eta}\eta^{-p})^{-j} 2|D^1 r_\delta|_\infty |F| + (\tilde{\eta}\eta^{-p})^{-j}|D^2 r_\delta|_\infty \eta^j |\Psi^{(1)}_{\eta j}(z)(\tilde{z} - z)||F|]|z_1| \cdots |z_p| \\
&\le \frac{\epsilon}{c(\tilde{\eta}) + 1}|z_1| \cdots |z_p|.
\end{aligned}$$

As $r_\delta|_{\{x \in E; |x_s| < \delta\}}$ is C^k-smooth, $k \ge 2$, $\Phi_i : P \to E$ is C^1-smooth and $D\Phi_j(z)\tilde{z} = \Psi^{(1)}_{\eta j}(z)\tilde{z}$ for $z \in P$ and $\tilde{z} \in \Lambda$. For any $\epsilon > 0$, there exists $\delta > 0$ so that when $\tilde{z} \in P$ and $|\tilde{z} - z| < \delta$, then for $0 \le j \le J_0$ we have

$$|Dr_\delta(\Phi_j(\tilde{z})) - Dr_\delta(\Phi_j(z)) - D^2 r_\delta(\Phi_j(z))\Psi^{(1)}_{\eta j}(z)(\tilde{z} - z)| < \frac{\tilde{\eta}^j \eta^{-pj}}{|F| + 1} \frac{\epsilon}{c(\tilde{\eta}) + 1},$$

and hence

$$\begin{aligned}
&\tilde{\eta}^{-j}|D^1 r_\delta(\Phi_j(\tilde{z}))F_j(z_1, \dots, z_p) - D^1 r_\delta(\Phi_j(z))F_j(z_1, \dots, z_p) \\
&\quad - D^2 r_\delta(\Phi_j(z))(\Psi^{(1)}_{\eta j}(z)(\tilde{z} - z), F_j(z_1, \dots, z_p))| \\
&< \tilde{\eta}^{-j} \frac{\tilde{\eta}^j \eta^{-pj}}{|F| + 1} \frac{\epsilon}{c(\tilde{\eta}) + 1} \eta^{pj}|F||z_1| \cdots |z_p| \\
&\le \frac{\epsilon}{c(\tilde{\eta}) + 1}|z_1| \cdots |z_p|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&|KA^{(1)}_{r_\delta}(\Phi(\tilde z))F - KA^{(1)}_{r_\delta}(\Phi(z))F - KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)(\tilde z - z), F)|\\
&\le c(\tilde\eta) \sup_{z_i\in\Lambda,|z_i|\le 1,1\le i\le p,j\ge 0} \tilde\eta^{-j}|[D^1 r_\delta(\Phi_j(\tilde z))F_j(z_1,\dots,z_p)\\
&\qquad - D^1 r_\delta(\Phi_j(z))F_j(z_1,\dots,z_p) - D^2 r_\delta(\Phi_j(z))(\Psi^{(1)}_{\eta j}(z)(\tilde z - z), F_j(z_1,\dots,z_p))]|\\
&< c(\tilde\eta)\frac{\epsilon}{c(\tilde\eta)+1} \le \epsilon.
\end{aligned}$$

This proves the differentiability.

We now prove that the map

$$P\times N \ni (z,F) \mapsto KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)\cdot, F) \in \mathcal{L}(\Lambda, \mathcal{L}^{(p)}(\Lambda, E_{\tilde\eta})) = \mathcal{L}^{(p+1)}(\Lambda, E_{\tilde\eta})$$

is continuous. Fix $(z,F)\in P\times N$. Then for any $(\tilde z,\tilde F)\in P\times N$, we have

$$\begin{aligned}
&|KA^{(2)}_{r_\delta}(\Phi(\tilde z))(\Psi^{(1)}_\eta(\tilde z)\cdot,\tilde F) - KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)\cdot, F)|\\
&\le |KA^{(2)}_{r_\delta}(\Phi(\tilde z))(\Psi^{(1)}_\eta(\tilde z)\cdot,\tilde F) - KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)\cdot,\tilde F)|\\
&\quad + |KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)\cdot,\tilde F) - KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)\cdot, F)|
\end{aligned}$$

and

$$\begin{aligned}
&|KA^{(2)}_{r_\delta}(\Phi(\tilde z))(\Psi^{(1)}_\eta(\tilde z)\cdot,\tilde F) - KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)\cdot,\tilde F)|\\
&= \sup_{z_i\in\Lambda,|z_i|\le 1,1\le i\le p+1} |KA^{(2)}_{r_\delta}(\Phi(\tilde z))(\Psi^{(1)}_\eta(\tilde z)z_{p+1},\tilde F(z_1,\dots,z_p))\\
&\qquad\qquad - KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)z_{p+1},\tilde F(z_1,\dots,z_p))|_{E_{\tilde\eta}}\\
&= \sup_{z_i\in\Lambda,|z_i|\le 1,1\le i\le p+1} |K_{\tilde\eta}[A^{(2)}_{r_\delta}(\Phi(\tilde z))(\Psi^{(1)}_\eta(\tilde z)z_{p+1},\tilde F(z_1,\dots,z_p))\\
&\qquad\qquad - A^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)z_{p+1},\tilde F(z_1,\dots,z_p))]|_{E_{\tilde\eta}}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&|A^{(2)}_{r_\delta}(\Phi(\tilde z))(\Psi^{(1)}_\eta(\tilde z)z_{p+1},\tilde F(z_1,\dots,z_p)) - A^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)z_{p+1},\tilde F(z_1,\dots,z_p))|_{E_{\tilde\eta}}\\
&= \sup_{j\in\mathbb{N}} \tilde\eta^{-j}|D^2 r_\delta(\Phi_j(\tilde z))(\Psi^{(1)}_{\eta j}(\tilde z)z_{p+1},\tilde F_j(z_1,\dots,z_p))\\
&\qquad - D^2 r_\delta(\Phi_j(z))(\Psi^{(1)}_{\eta j}(z)z_{p+1},\tilde F_j(z_1,\dots,z_p))|.
\end{aligned}$$

Note that for any $\eta^*\in(\eta,\bar\eta]$, the mapping $j_{\eta^*\eta}\circ\Psi^{(1)}_\eta : P\to E_{\eta^*}$ is continuous. Fix $\eta^*\in(\eta,\frac{\bar\eta}{\eta^p})$. There exists $\delta_1>0$ so that if $\tilde z\in P$ and $|\tilde z - z|<\delta_1$, then

$$|j_{\eta^*\eta}\circ\Psi^{(1)}_\eta(\tilde z) - j_{\eta^*\eta}\circ\Psi^{(1)}_\eta(z)|\le 1.$$

Therefore, $\eta^{*-j}|\Psi^{(1)}_{\eta j}(\tilde z) - \Psi^{(1)}_{\eta j}(z)|\le 1$ for all $j\in\mathbb{N}$. In particular, $|\Psi^{(1)}_{\eta j}(\tilde z) - \Psi^{(1)}_{\eta j}(z)|\le \eta^{*j}$ for all $j\in\mathbb{N}$.

Find an integer $J_0\ge 0$ so that if $j\ge J_0$, then

$$|D^2 r_\delta|_\infty\left(\frac{\tilde\eta}{\eta^p}\right)^{-j}[2\eta^j|\Psi^{(1)}_\eta(z)| + (\eta^*)^j] < \frac{\epsilon}{2(c(\eta^*)+1)(|F|+1)}.$$

Therefore, for $j \geq J_0$, we have

$$\begin{aligned}
&\tilde{\eta}^{-j}|D^2 r_\delta(\Phi_j(\tilde{z}))(\Psi_\eta^{(1)}(\tilde{z})z_{p+1}, \tilde{F}_j(z_1,\dots,z_p)) - D^2 r_\delta(\Phi_j(z))(\Psi_{\eta j}^{(1)}(z)z_{p+1}, \tilde{F}_j(z_1,\dots,z_p))| \\
&\leq |D^2 r_\delta|_\infty \tilde{\eta}^{-j}[2|\Psi_{\eta j}^{(1)}(z)| + \eta^{*j}]\eta^{pj}|\tilde{F}||z_1|\cdots|z_p||z_{p+1}| \\
&\leq |D^2 r_\delta|_\infty \left(\frac{\tilde{\eta}}{\eta^p}\right)^{-j} [2\eta^j|\Psi_\eta^{(1)}(z)| + \eta^{*j}]|\tilde{F}||z_1|\cdots|z_{p+1}|.
\end{aligned}$$

For $0 \leq j \leq J_0$, as $\Phi_j = e_j\Phi$ and $\Psi_{\eta j}^{(1)} = e_j j_{\eta^*\eta}\Psi_{\eta j}^{(1)}$ are continuous, we can find $\delta_2 > 0$ so that when $\tilde{z} \in P$ and $|\tilde{z} - z| < \delta_2$, we have

$$\begin{aligned}
&\tilde{\eta}^{-j}|D^2 r_\delta(\Phi_j(\tilde{z}))(\Psi_{\eta j}^{(1)}(\tilde{z})z_{p+1}, \tilde{F}_j(z_1,\dots,z_p)) - D^2 r_\delta(\Phi_j(z))(\Psi_{\eta j}^{(1)}(z)z_{p+1}, \tilde{F}_j(z_1,\dots,z_p))| \\
&\quad < \frac{\epsilon}{2(c(\tilde{\eta})+1)(|F|+1)}|\tilde{F}||z_1|\cdots|z_{p+1}|.
\end{aligned}$$

Therefore, if $|\tilde{F} - F| \leq 1$ and $|\tilde{z} - z| < \min\{\delta_1, \delta_2\}$, we have

$$\begin{aligned}
&|KA_{r_\delta}^{(2)}(\Phi(\tilde{z}))(\Psi_\eta^{(1)}(\tilde{z})\cdot, \tilde{F}) - KA_{r_\delta}^{(2)}(\Phi(z))(\Psi_\eta^{(1)}(z)\cdot, \tilde{F})| \\
&\quad \leq c(\tilde{\eta})\frac{\epsilon}{2(c(\tilde{\eta})+1)} < \frac{\epsilon}{2}.
\end{aligned}$$

In a similar fashion, we get

$$\begin{aligned}
&|KA_{r_\delta}^{(1)}(\Phi(z))(\Psi_\eta^{(1)}(z), \tilde{F})_K A_{r_\delta}^{(2)}(\Phi(z))(\Psi_\eta^{(1)}(z), F)| \\
&= \sup_{z_i\in\Lambda, |z_i|\leq 1, 1\leq i\leq p+1} |KA_{r_\delta}^{(2)}(\Psi_\eta^{(1)}(z)z_{p+1}, (\tilde{F} - F)(z_1,\dots,z_p))|_{E_{\tilde{\eta}}} \\
&\leq c(\tilde{\eta}) \sup_{z_i\in\Lambda, |z_i|\leq 1, 1\leq i\leq p+1, j\geq 0} \tilde{\eta}^{-j}|D^2 r_\delta|_\infty \eta^{-j}|\Psi_\eta^{(1)}(z)|\eta^{-pj}|\tilde{F} - F||z_1|\cdots|z_{p+1}| \\
&\leq c(\tilde{\eta})|D^2 r_\delta|_\infty |\Psi_\eta^{(1)}(z)||\tilde{F} - F|.
\end{aligned}$$

Therefore, if $|\tilde{z} - z| < \min\{\delta_1, \delta_2\}$ and if $|\tilde{F} - F| < \min\{1, \frac{\epsilon}{2c(\tilde{\eta})|D^2 r_\delta|_\infty |\Psi^{(1)}(z)|+1}\}$, then $|KA_{r_\delta}^{(1)}(\Phi(\tilde{z})\cdot, \tilde{F}) - KA_{r_\delta}^{(1)}(\Phi(z)\cdot, F)| < \epsilon$. This completes the proof of the required continuity.

For the sake of later reference, let us summarize the main idea of the arguments involved in this step. To estimate

$$|KA_{r_\delta}^{(1)}(\Phi(\tilde{z}))F - KA_{r_\delta}^{(1)}(\Phi(z))F - KA_{r_\delta}^{(2)}(\Phi(z))(\Psi_\eta^{(1)}(z)(\tilde{z} - z), F)|$$

in the proof of the differentiability of the mapping $P \ni z \mapsto KA_{r_\delta}^{(1)}(\Phi(z))F \in \mathcal{L}^{(p)}(\Lambda, E_{\tilde{\eta}})$, we used the definition of the operator norm for multilinear operators $KA_{r_\delta}^{(1)}(\Phi(z))F$ and the definition of the norm in $E_{\tilde{\eta}}$ and were led to the estimation of the expression

$$\begin{aligned}
&\tilde{\eta}^{-j}|[D^1 r_\delta(\Phi_j(\tilde{z}))F_j(z_1,\dots,z_p) - D^1 r_\delta(\Phi_j(z))F_j(z_1,\dots,z_p) \\
&\qquad - D^2 r_\delta(\Phi_j(z))(\Psi_{\eta j}^{(1)}(z)(\tilde{z} - z), F_j(z_1,\dots,z_p))|
\end{aligned}$$

for each given nonnegative integer j. The above term can be made arbitrarily small if j is sufficiently large, thanks to the choice of $\tilde{\eta} > \eta^{p+1}$ (the essential gradient of the proof). When j is restricted to a finite set, the smallness of the above expression

follows from the continuity of the involved operators and mappings. Similar arguments were used to estimate

$$|KA^{(2)}_{r_\delta}(\Phi(\tilde z))(\Psi^{(1)}_\eta(\tilde z)\cdot,\tilde F)-KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)\cdot,F)|$$

in the proof of the continuity of the map

$$P\times N\ni(z,F)\mapsto KA^{(2)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)\cdot,F)\in\mathcal{L}(\Lambda,\mathcal{L}^{(p)}(\Lambda,E_{\tilde\eta})).$$

Step 6. Let $2\le l\le p, 1\le r_i<l$ with $r_1+\cdots+r_l=p$. For any integer $j\ge 0$, $z\in\Lambda$, and $\hat z_{r_i}\in\Lambda^{r_i}$, let

$$\Psi^{(r_i)}_{\eta j}(z)\hat z_{r_i}=e_j(\Psi^{(r_i)}_\eta(z)\hat z_{r_i}).$$

Then for $z,\tilde z\in\Lambda$ we have

$$\begin{aligned}
&\Big|KA^{(l)}_{r_\delta}(\Phi(\tilde z))(\Psi^{(r_1)}_\eta(\tilde z),\dots,\Psi^{(r_l)}_\eta(\tilde z))-KA^{(l)}_{r_\delta}(\Phi(z))(\Psi^{(r_1)}_\eta(z),\dots,\Psi^{(r_l)}_\eta(z))\\
&\quad-\sum_{k=1}^{l}KA^{(l)}_{r_\delta}(\Phi(z))(\Psi^{(r_1)}_\eta(z),\dots,\Psi^{(r_k+1)}_\eta(z)(\tilde z-z),\dots,\Psi^{(r_l)}_\eta(z))\\
&\quad-KA^{(l+1)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z)(\tilde z-z),\Psi^{(r_1)}_\eta(z),\dots,\Psi^{(r_l)}_\eta(z))\Big|\\
&\le c(\tilde\eta)\sup_{\hat z_{r_i}\in\Lambda^{r_i},|z_{r_i}|\le 1,1\le i\le p,j\ge 0}\tilde\eta^{-j}\Big|D^l r_\delta(\Phi_j(\tilde z))(\Psi^{(r_1)}_{\eta j}(\tilde z)\hat z_{r_1},\dots,\Psi^{(r_l)}_{\eta j}(\tilde z)\hat z_{r_l})\\
&\quad-D^l r_\delta(\Phi_j(z))(\Psi^{(r_1)}_{\eta j}(z)\hat z_{r_1},\dots,\Psi^{(r_l)}_{\eta j}(z)\hat z_{r_l})\\
&\quad-\sum_{k=1}^{l}D^l r_\delta(\Phi_j(z))(\Psi^{r_1}_{\eta j}(z)\hat z_{r_1}\dots,\Psi^{(r_k+1)}_{\eta j}(z)(\tilde z-z,\hat z_{r_k}),\dots,\Psi^{(r_l)}_{\eta j}(z)z_{r_l})\\
&\quad-D^{l+1}r_\delta(\Phi_j(z))(\Psi^{(1)}_{\eta j}(z)(\tilde z-z),\Psi^{(r_1)}_{\eta j}(z)\hat z_{r_1},\dots,\Psi^{(r_l)}_{\eta j}(z)\hat z_{r_l})\Big|.
\end{aligned}$$

Now we can use the fact that $|D^l r_\delta|_\infty<\infty$ for $1\le l\le p$, and the induction hypothesis implies that the mapping

$$P\ni z\mapsto\Psi^{(r_i)}_{\eta j}(z)\in\mathcal{L}^{(r_i)}(\Lambda,E_{\eta^{r_i}})$$

is differentiable, and we apply an argument similar to that for the first part of Step 5 to show that for any $2\le l\le p$, $1\le r_i<l$ with $r_1+\cdots+r_l=p$, the map $P\ni z\mapsto KA^{(l)}_{r_\delta}(\Phi(z))(\Psi^{(r_1)}_\eta(z),\dots,\Psi^{(r_l)}_\eta(z))\in\mathcal{L}^{(p)}(\Lambda,E_{\tilde\eta})$ is differentiable and the derivative is given by

$$\begin{aligned}
&\sum_{j=1}^{l}KA^{(l)}_{r_\delta}(\Phi(z))(\Psi^{(r_1)}_\eta(z),\dots,\Psi^{(r_j+1)}_\eta(z),\dots,\Psi^{(r_l)}_\eta(z))\\
&\quad+KA^{(l+1)}_{r_\delta}(\Phi(z))(\Psi^{(1)}_\eta(z),\Psi^{(r_1)}_\eta(z),\dots,\Psi^{(r_l)}_\eta(z)).
\end{aligned}$$

The continuity of the above derivative, with respect to $z \in P$, can also be verified by using an argument similar to that for the second part of Step 5 and by noting that the induction hypothesis implies that the mapping

$$P \ni z \mapsto \Psi_{\eta j}^{(r_i+1)}(z) \in \mathcal{L}^{(r_i+1)}(\Lambda, E_{\eta^{r_i}})$$

is continuous.

Step 7. Let $\hat{\eta}$ be given so that $\hat{\eta} \in (\tilde{\eta}, \bar{\eta}]$. Define the continuous linear injective map $J : X \to X_1 = \mathcal{L}^{(p)}(\Lambda, E_{\hat{\eta}})$ by

$$J(L)(z_1, \ldots, z_p) = j_{\hat{\eta}\tilde{\eta}} L(z_1, \ldots, z_p), \qquad z_1, \ldots, z_p \in \Lambda, \quad L \in X.$$

Then

$$JH_0(F, z) = j_{\hat{\eta}\tilde{\eta}} K A_{r_\delta}^{(1)}(\Phi(z))F + j_{\hat{\eta}\eta} H_p(z), \qquad z \in P, \quad F \in \mathcal{L}^{(p)}(\Lambda, E_{\eta^p}).$$

Let $A : P \to \mathcal{L}(X, X_1)$ be given by

$$(A(z)F)(z_1, \ldots, z_p) = j_{\hat{\eta}\tilde{\eta}} K \circ A_{r_\delta}(\Phi(z))F(z_1, \ldots, z_p), \qquad z \in P, \ F \in X, \ x_1, \ldots, z_p \in \Lambda.$$

Again, we can use arguments similar to those in Step 5 (see the remarks at the end of Step 5) to show that A is continuous. Moreover, we have

$$JH_0(\tilde{F}, z) - JH_0(F, z) = A(z)(\tilde{F} - F), \qquad z \in P, \quad \tilde{F}, F \in N.$$

Note that for any $\eta^* \geq \eta$, $KA_{r_\delta}^{(1)}(\Phi(z))$ induces a bounded linear map from $\mathcal{L}^{(p)}(\Lambda, E_{\eta^*})$ into itself by

$$Q_{\eta^*}(L)(z_1, \ldots, z_p) = K_{\eta^*} A_{r_\delta \eta^* \eta^*}^{(1)}(\Phi(z)) L(z_1, \ldots, z_p)$$

and

$$|Q_{\eta^*}| \leq c(\eta^*)\lambda(\delta).$$

Define $H^{(1)} : P \to \mathcal{L}(X, X)$ and $H_1^{(1)} : P \to \mathcal{L}(X_1, X_1)$ by

$$H^{(1)}(z) = Q_{\tilde{\eta}}, \quad H_1^{(1)}(z) = Q_{\hat{\eta}}, \qquad z \in P.$$

Clearly, we have for $F \in X$ the following:

$$\begin{aligned} A(z)F &= j_{\hat{\eta}\tilde{\eta}} K A_{r_\delta}^{(1)}(\Phi(z))F \\ &= j_{\hat{\eta}\tilde{\eta}} Q_{\tilde{\eta}} F = JH_1^{(1)}(z)F \\ &= Q_{\hat{\eta}} j_{\hat{\eta}\tilde{\eta}} F = H_1^{(1)}(z)JF \end{aligned}$$

and

$$|H^{(1)}(z)| \leq c(\tilde{\eta})\lambda(\delta) \leq \kappa, \qquad |H_1^{(1)}(z)| \leq c(\hat{\eta})\lambda(\delta) \leq \kappa.$$

Moreover, the mapping

$$P \ni z \mapsto J \circ H^{(1)}(z) = j_{\hat{\eta}\tilde{\eta}} \circ Q_{\tilde{\eta}} = J_{\hat{\eta}\tilde{\eta}} K_{\tilde{\eta}} A_{r_\delta \tilde{\eta}\tilde{\eta}}^{(1)}(\Phi(z)) = A \in \mathcal{L}(X, X_1)$$

is continuous. Therefore, by Lemma 5.2, the map $j_{\hat{\eta}\eta}\circ\Psi_\eta^{(p)} = j_{\hat{\eta}\tilde{\eta}}\circ j_{\tilde{\eta}\eta}\circ\Psi_\eta^{(p)} : P \to X_1$ is C^1-smooth and

$$D(j_{\hat{\eta}\eta}\circ\Psi_\eta^{(p)})(z) = KA_{r_\delta}^{(1)}(\Phi(z))D(j_{\hat{\eta}\tilde{\eta}}\circ\Psi_\eta^{(p)}) + j_{\hat{\eta}\tilde{\eta}}\circ D_2H_0(\Psi_\eta^{(p)}, z), \qquad z \in P.$$

Step 8. We now prove that the mapping $j_{\hat{\eta}\eta}\circ\Phi : P \to E_{\hat{\eta}}$ is C^{p+1}-smooth. Indeed, as $\hat{\eta} > \eta^{p+1} > \eta^p$, $j_{\hat{\eta}\eta}\circ\Phi : P \to E_{\hat{\eta}}$ is C^p-smooth and

$$D^p(j_{\hat{\eta}\eta}\circ\Phi) = j_{\hat{\eta}\eta}\circ\Psi_\eta^{(p)}.$$

Since $j_{\hat{\eta}\eta}\circ\Psi_\eta^{(p)}$ is C^1-smooth, we conclude that $j_{\hat{\eta}\eta}\circ\Phi$ is C^{p+1}-smooth and $D^{p+1}(j_{\hat{\eta}\eta}\circ\Phi) = D(j_{\hat{\eta}\eta}\circ\Psi_\eta^{(p)})$. Let $H_{p+1}(z) = D_2H_0(\Psi_\eta^p(z), z)$ and let $\Psi_\eta^{(p+1)}(z)$ be the unique fixed point of the contraction

$$\mathcal{L}^{(p+1)}(\Lambda, E_{\eta^{p+1}}) \ni F \mapsto K_{\eta^{p+1}}A^{(1)}_{r_\delta\eta^{p+1}\eta^{p+1}}(\Phi(z))F + H_{p+1}(z) \in \mathcal{L}^{(p+1)}(\Lambda, E_{\eta^{p+1}});$$

then $D^{p+1}(j_{\hat{\eta}\eta}\circ\Phi) = j_{\hat{\eta}\eta}\Psi_\eta^{(p+1)}$. This proves all conclusions in the case of $p+1$.

Therefore, we have proved that for a fixed $\tilde{\eta} > \eta^k$ the mapping $j_{\tilde{\eta}\eta}\circ\Phi : P \to E_{\eta^k}$ is C^k-smooth, and hence $\chi_{\delta_{\tilde{\eta}}}|_P = j_{\tilde{\eta}\eta}\Phi$ is C^k smooth. Consequently, $w_{\delta_{\tilde{\eta}}}|_P = P_u \circ e_0\chi_{\delta_{\tilde{\eta}}}|P$ is C^k-smooth. □

Similarly, we have the following center-unstable manifold theorem.

THEOREM 5.3. *Let $f : U \to E$ be a C^1-map on an open subset U of a Banach space E over $\mathbb{R}$, with a fixed point p. Let $L = Df(p)$ and assume that E has the following decomposition:*

$$E = E_s \oplus E_c \oplus E_u,$$

where E_s is a closed subspace, E_c and E_u are finite-dimensional, $L(E_s) \subset E_s$, $L(E_c) \subset E_c$, and $L(E_u) \subset E_u$. We further assume that

$$\sigma_s = \sigma(L|_{E_s} : E_s \to E_s) \text{ is contained in a compact subset of } \{z \in \mathbb{C} : |z| < 1\}$$

and

$$\sigma_c = \sigma(L|_{E_c} : E_c \to E_c) \subset S^1_{\mathbb{C}},$$
$$\sigma_u = \sigma(L|_{E_u} : E_u \to E_u) \subset \{z \in \mathbb{C} : |z| > 1\}.$$

Let $E_{cu} = E_u \oplus E_c$. Then

(i) *there exist open neighborhoods N_{cu} of 0 in E_{cu}, N_s of 0 in E_s, N of p in U, and a C^1-map $w : N_{cu} \to E_s$ with $w(0) = 0$, $Dw(0) = 0$, and $w(N_{cu}) \subset N_s$ so that the shifted graph $W = p + \{z + w(z) : z \in N_{cu}\}$ satisfies $f(W \cap N) \subset W$ and $\{x \in E;$ there exists a trajectory $(x_n)_{-\infty}^0$ of f in $p + N_{cu} + N_s$ with $x_0 = x\} \subset W$;*

(ii) *if f is C^k-smooth for an integer $k \geq 2$, then so is w.*

We can now state the following smoothness theorem for center manifolds in general Banach spaces.

THEOREM 5.4. *Let $f : U \to E$ be a C^1-map on an open subset U of a Banach space E over $\mathbb{R}$, with a fixed point p. Let $L = Df(p)$ and assume that E has the following decomposition:*

$$E = E_s \oplus E_c \oplus E_u,$$

where E_s is a closed subspace, E_c and E_u are finite-dimensional, $L(E_s) \subset E_s$, $L(E_c) \subset E_c$, and $L(E_u) \subset E_u$. We further assume that

$$\sigma_s = \sigma(L|_{E_s} : E_s \to E_s) \text{ is contained in a compact subset of } \{z \in \mathbb{C} : |z| < 1\}$$

and

$$\begin{aligned}\sigma_c =& \sigma(L|_{E_c} : E_c \to E_c) \subset S^1_{\mathbb{C}},\\ \sigma_u =& \sigma(L|_{E_u} : E_u \to E_u) \subset \{z \in \mathbb{C} : |z| > 1\}.\end{aligned}$$

Let $E_{su} = E_s \oplus E_c$. Then

(i) *there exist open neighborhoods N_c of 0 in E_c, N_{su} of 0 in E_{su}, N of p in U, and a C^1-map $w : N_c \to E_{su}$ with $w(0) = 0$, $Dw(0) = 0$, and $w(N_c) \subset N_{su}$ so that the shifted graph $W = p + \{z + w(z) : z \in N_c\}$ satisfies $f(W \cap N) \subset W$, and if there exists $(x_n)_{-\infty}^{\infty}$ such that $x_n = f(x_{n-1})$ and $x_n \in p + N_c + N_{su}$ for every integer n, then $x_0 \in W$;*

(ii) *if f is C^k-smooth for an integer $k \geq 2$, then so is w.*

Proof. Without loss of generality, we may assume $p = 0$. By Theorem 5.1, there exist convex open neighborhoods $\widetilde{N}_{cs}$ of 0 in $E_c + E_s$, $\widetilde{N}_u$ of 0 in E_u, $\widetilde{N}$ of 0 in U, and a C^k-map ($k = 1$ in case of (i) and $k \geq 2$ in case of (ii)) $\widetilde{w}_{cs} : \widetilde{N}_{cs} \to E_u$ with

$$\begin{aligned}&\widetilde{w}_{cs}(0) = 0, \qquad D\widetilde{w}_{cs}(0) = 0;\\ &\widetilde{w}_{cs}(\widetilde{N}_{cs}) \subset \widetilde{N}_u,\end{aligned}$$

and such that the graph

$$\widetilde{W}_{cs} = \{z_{cs} + \widetilde{w}_{cs}(z_{cs}) : z_{cs} \in \widetilde{N}_{cs}\}$$

satisfies

$$f(\widetilde{W}_{cs} \cap \widetilde{N}) \subset \widetilde{W}_{cs}$$

and

$$\cap_{n=0}^{\infty} f^{-n}(\widetilde{N}_{cs} + \widetilde{N}_u) \subset \widetilde{W}_{cs}. \tag{5.3}$$

By Theorem 5.3, there exist open neighborhoods $\widehat{N}_{cu}$ of 0 in $E_c \oplus E_u$, $\widehat{N}_s$ of 0 in E_s, $\widehat{N}$ of 0 in U, and a C^k-map ($k = 1$ in case (i) and $k \geq 2$ in case (ii)) $\widehat{w}_{cu} : \widehat{N}_{cu} \to E_s$ with

$$\begin{aligned}&\widehat{w}_{cu}(0) = 0, \qquad D\widehat{w}_{cu}(0) = 0;\\ &\widehat{w}_{cu}(\widehat{N}_{cu}) \subset \widehat{N}_s,\end{aligned}$$

and the graph

$$\widehat{W}_{cu} = \{z_{cu} + \widehat{w}_{cu}(z_{cu}) : z_{cu} \in \widehat{N}_{cu}\}$$

satisfies

$$f(\widehat{W}_{cu} \cap \widehat{N}) \subset \widehat{W}_{cu}$$

and

$$\begin{aligned}&z \in \widehat{W}_{cu} \text{ if there exists } \{z_n\}_{n=-\infty}^{0} \subset \widehat{N}_{cu} + \widehat{N}_s\\ &\text{such that } z_{n+1} = f(z_n) \text{ for } n \leq -1 \text{ and that } z_0 = z.\end{aligned} \tag{5.4}$$

Choose open neighborhoods N_c^* of 0 in E_c, N_s^* of 0 in E_s, N_c^* of 0 in E_c, N^* of 0 in E such that

$$\begin{cases} N^* \subset \widehat{N} \cap \widetilde{N}; \\ N_c^* + N_s^* \subset \widetilde{N}_{cs}; \\ N_c^* + N_u^* \subset \widehat{N}_{cu}; \\ z_c \in N_c^*, z_s \in N_s^* \text{ if } z \in f(N^*); \\ z_c \in N_c^*, z_u \in N_u^* \text{ if } z \in f(N^*); \\ \widetilde{w}_{cs}(z_c + z_s) \in N_u^* \text{ if } z_c \in N_c^* \text{ and } z_s \in N_s^*. \end{cases}$$

Define

$$W_{cs}^* = \{z_{cs} + \widetilde{w}_{cs}(z_{cs}) : z_{cs} = z_c + z_s \in N_c^* + N_s^*\},$$
$$W_{cu}^* = \{z_{cu} + \widehat{w}_{cu}(z_{cu}) : z_{cu} = z_c + z_u \in N_c^* + N_u^*\},$$

and

$$W^* = W_{cs}^* \cap W_{cu}^*.$$

For $z \in W^*$, we have

$$\begin{aligned} z &= z_c + z_s + \widetilde{w}_{cs}(z_c + z_s) \\ &= z_c + z_u + \widehat{w}_{cu}(z_c + z_u) \end{aligned}$$

with $z_c \in N_c^*, z_s \in N_s^*$, and $z_u \in N_u^*$. Therefore,

$$z_s = \widehat{w}_{cu}(z_c + z_u) = \widehat{w}_{cu}(z_c + \widetilde{w}_{cs}(z_c + z_s)).$$

Consider the equation

$$z_s = \widehat{w}_{cu}(z_c + \widetilde{w}_{cs}(z_c + z_s)). \tag{5.5}$$

As both $\widehat{w}_{cu}$ and $\widetilde{w}_{cs}$ are C^k-smooth and $D\widehat{w}_{cu}(0) = 0, D\widetilde{w}_{cs}(0) = 0$, the implicit function theorem implies that there are open neighborhoods N_c of 0 in N_c^* and N_s of 0 in N_s^* and a C^k-map $w_s : N_c \to N_s$ such that for every $z_c \in N_s$ equation (5.5) has the unique solution $z_s = w_s(z_c)$. It is easy to verify that $w_s(0) = 0$ and $Dw_s(0) = 0$.

We now define $w_c : N_c \to E_s \oplus E_u$ by

$$w_c(z_c) = w_s(z_c) + \widetilde{w}_{cs}(z_c + w_s(z_c)), \qquad z_c \in N_c.$$

Clearly, w_c is C^k-smooth, $w_c(0) = 0$, $Dw_c(0) = 0$, and

$$w_c(N_c) \subset N_s + N_u$$

with

$$N_u = N_u^*.$$

Let

$$W_c = \{z_c + w_c(z_c) : z_c \in N_c\}.$$

We prove that if there exists $\{z_n\}_{n=-\infty}^{\infty} \subset N_c + N_s + N_u$ such that $z_{n+1} = f(z_n)$ for $n \in \mathbf{Z}$, then $z = z_0 \in W_c$. In fact, (5.3) and (5.4) imply that $z \in \widehat{W}_{cu} \cap \widetilde{W}_{cs}$. As $z_c \in N_c \subset N_c^*$, $z_s \in N_s \subset N_s^*$, and $z_u \in N_u = N_u^*$, we have $z \in W^*$ and

$$z_0 = z_c + z_s + \widetilde{w}_{cs}(z_c + z_s) = z_c + z_u + \widehat{w}_{cu}(z_c + z_u),$$

from which it follows that

$$z_s = \widehat{w}_{cu}(z_c + \widetilde{w}_{cs}(z_c + z_s)), \qquad z_s \in N_s, \quad z_c \in N_c.$$

Therefore, we must have $z_s = w_s(z_c)$ and $z_u = \widetilde{w}_{cs}(z_c + w_s(z_c))$. This shows that $z \in W_c$.

Other properties in Theorem 5.4 are straightforward consequences of Theorems 5.1 and 5.3. □

6. Center manifolds for nonlinear FDEs in Banach spaces. We now start to consider semilinear FDEs

$$\dot{u}(t) = A_T u(t) + L(u_t) + F(u_t), \tag{6.1}$$

where we assume A_T, L are as in the previous sections, and, in particular, that (H1)–(H3) are satisfied. We also assume that $F : V_1 \to X$ is a C^k-mapping ($k \geq 1$) from a neighborhood V_1 of $0 \in C$ into X with $F(0) = 0$ and $DF(0) = 0$.

Fix $\omega > r$. Using the arguments of Fitzgibbon [6] (see also Theorems 2.1 and 2.2 in Chapter 2 of Wu [21]), we can find an open neighborhood $V_2 \subset V_1$ of 0 in C such that for any $\phi \in V_2$ there exists a unique continuous function $u^\phi : [-r, \omega] \to X$ such that $u_0^\phi = \phi$ and

$$u^\phi(t) = T(t)\phi(0) + \int_0^t T(t-s)[L(u_s^\phi) + F(u_s^\phi)]ds$$

for $t \in [0, \omega]$. Define $\tilde{f} : V_2 \to C$ by

$$\tilde{f}(\phi) = u_\omega^\phi \quad \text{for } \phi \in V_2.$$

As $\omega > r$, we can show that $\tilde{f}$ is compact (using the argument in Travis and Webb [18]; see also Theorem 1.8 of Chapter 2 of Wu [21]). The next lemma shows that there exists an open neighborhood $V \subset V_2$ of 0 in C such that $f = \tilde{f}|_V : V \to C$ is C^k-smooth and

$$Df(0) = U(\omega) : C \to C.$$

LEMMA 6.1. *There exists an open neighborhood $V \subset V_2$ of 0 in C such that for each $t \in [0, \omega]$, u_t^ϕ is C^k-smooth with respect to $\phi \in V$. Moreover, for each $\psi \in C$, $D_\phi u^\phi(t)\psi$ satisfies the linear variational equation*

$$\begin{cases} v(t) = T(t)\psi(0) + \int_0^t T(t-s)[L(v_s) + DF(u_s^\phi)v_s]ds, \\ v_0 = \psi. \end{cases} \tag{6.2}$$

In particular, $Df(0) = U(\omega)$.

Proof. We are going to apply the same argument as that for Theorem 4.1 in Hale [8] based on [8, Lemma 4.2, p. 46]. Let $\hat{F}(\phi) = L(\phi) + F(\phi)$. Fix $\chi \in V_2$. There exist $M > 0$, $\delta > 0$, and $N > 0$ such that

$$\begin{cases} \|T(t)\| \leq M \quad \text{for } t \in [0, 1]; \\ \overline{B_\delta(\chi)} \subset V_2 \quad \text{with } B_\delta(\chi) = \{\psi \in C : ||\psi - \chi|| < \delta\}; \\ |\hat{F}(\psi)| \leq N, \ |D\hat{F}(\psi)| \leq N \text{ for } \psi \in \overline{B_\delta(\chi)}. \end{cases}$$

Now choose $\epsilon \in (0,1)$ and $\beta \in (0,1)$ so that

$$\begin{cases} \beta < \frac{\delta}{2}; \\ \sup_{\theta,\theta' \in [-r,0], |\theta-\theta'| \le \epsilon} |\chi(\theta) - \chi(\theta')| < \frac{\delta}{8}; \\ \sup_{t \in [0,\epsilon]} |T(t)\chi(0) - \chi(0)| < \frac{\delta}{8}; \\ \epsilon < \frac{\beta}{MN}. \end{cases}$$

Let

$$K(\epsilon,\beta) = \{y \in C([-r,\epsilon];X) : y_0 = 0,\ ||y_t|| \le \beta \text{ for } t \in [0,\epsilon]\}.$$

Clearly, $K(\epsilon,\beta)$ is a closed subset of the Banach space $C_0([-r,\epsilon]) = \{z \in C([-r,\epsilon];X) : z(s) = 0 \text{ for } s \in [-r,0]\}$ equipped with the supremum norm.

For each $\phi \in C$, define $\tilde{\phi} : [-r,\infty) \to X$ by $\tilde{\phi}_0 = \phi$ and $\tilde{\phi}(t) = T(t)\phi(0)$ for $t \ge 0$. Now, for fixed $\phi \in B_{\frac{\delta}{8(1+M)}}(\chi)$ define $A(\phi)$ on $K(\epsilon,\beta)$ by

$$A(\phi)y(t) = \begin{cases} \int_0^t T(t-s)\hat{F}(y_s + \tilde{\phi}_s)ds,\ y \in K(\epsilon,\beta), & t \in [0,\epsilon]; \\ 0, \quad t \in [-r,0]. \end{cases}$$

Clearly, $A(\phi)y \in C([-r,\epsilon];X)$. Moreover, since for $s \in [0,\epsilon]$, $||y_s|| \le \beta$, and

$$\begin{aligned} ||\tilde{\phi}_s - \chi|| &\le ||\tilde{\phi}_s - \tilde{\chi}_s|| + ||\tilde{\chi}_s - \chi|| \\ &\le ||\phi - \chi|| + \sup_{s\in[0,\epsilon]} ||T(s)||\phi(0) - \chi(0)| \\ &\quad + \sup_{\theta\in[-r,0],s\in[0,\epsilon],s+\theta\in[-r,0]} |\chi(\theta+s) - \chi(\theta)| \\ &\quad + \sup_{\theta\in[-r,0],s\in[0,\epsilon],s+\theta\ge 0} |T(s+\theta)\chi(0) - \chi(0)| \\ &\quad + \sup_{\theta\in[-r,0],s\in[0,\epsilon],s+\theta\ge 0} |\chi(\theta) - \chi(0)| \\ &\le (1+M)||\phi - \chi|| + \frac{\delta}{8} + \frac{\delta}{8} + \frac{\delta}{8} < \frac{\delta}{2}, \end{aligned}$$

we have

$$||y_s + \tilde{\phi}_s - \chi|| < \beta + \frac{\delta}{2} < \delta,$$

and hence

$$|\hat{F}(y_s + \tilde{\phi}_s)| \le N \quad \text{for } s \in [0,\epsilon].$$

This implies that

$$|A(\phi)y(t)| \le MN\epsilon < \beta \quad \text{for } t \in [0,\epsilon].$$

So, $A(\phi)y \in K(\epsilon,\beta)$ and $A(\phi)K(\epsilon,\beta) \subset K(\epsilon,\beta)$.

Moreover, using $||D\hat{F}(\psi)|| \le N$ for all $\psi \in \overline{B_\delta(\chi)}$, for $y, \hat{y} \in K(\epsilon,\beta)$ and $t \in [0,\epsilon]$ we have

$$\begin{aligned}
&|A(\phi)y(t) - A(\phi)\hat{y}(t)| \\
&\le \left| \int_0^t T(t-s)[\hat{F}(y_s + \tilde{\phi}_s) - \hat{F}(\hat{y}_s + \tilde{\phi}_s)ds] \right| \\
&\le MN\epsilon \sup_{s\in[0,t]} ||y_s - \hat{y}_s|| \\
&\le MN\epsilon \sup_{s\in[-r,\epsilon]} |y(s) - \hat{y}(s)| \\
&\le \beta \sup_{s\in[-r,\epsilon]} |y(s) - \hat{y}(s)|.
\end{aligned}$$

As $\beta < 1$, we conclude that for each $\phi \in \overline{B_{\frac{\delta}{8(1+M)}}(\chi)}$, the mapping $A(\phi) : K(\epsilon,\beta) \to K(\epsilon,\beta)$ is a contraction. By Lemma 4.2 of Hale [8], for each fixed $\phi \in \overline{B_{\frac{\delta}{8(1+M)}}(\chi)}$, $A(\phi)$ has a unique fixed point $y(\phi) \in K(\epsilon,\beta)$ which is continuous in ϕ.

Note that $\overline{B_{\frac{\delta}{8(1+M)}}(\chi)}$ is the closure of the open set $B_{\frac{\delta}{8(1+M)}}(\chi)$ and $A(\phi)y$ has a continuous kth derivative with respect to $(\phi, y) \in B_{\frac{\delta}{8(1+M)}}(\chi) \times K^0(\epsilon,\beta)$, where

$$K^0(\epsilon,\beta) = \{y \in K(\epsilon,\beta) : ||y_t|| < \beta \text{ for } t \in [0,\epsilon]\}$$

is open in $C_0([-r,\epsilon])$ and $K(\epsilon,\beta) = \overline{K^0(\epsilon,\beta)}$. Therefore, by Lemma 4.2 in Hale [8], $y(\phi)$ is C^k-smooth with respect to $\phi \in B_{\frac{\delta}{8(1+M)}}(\chi)$, and hence $u_t^\phi = \tilde{\phi}_t + (y(\phi))_t$ is C^k-smooth in $\phi \in B_{\frac{\delta}{8(1+M)}}(\chi)$ for each fixed $t \in [0,\epsilon]$. A standard continuation argument then leads to the C^k-smoothness of $u(\phi)$ with respect to ϕ for $t \in [0,\omega]$. The remaining part of the lemma can be easily verified. □

Let

$$\begin{aligned}
\Sigma_s &= \{\lambda \in \sigma_P(A_U) : \mathrm{Re}\lambda < 0\}, \\
\Sigma_u &= \{\lambda \in \sigma_P(A_U) : \mathrm{Re}\lambda > 0\}, \\
\Sigma_c &= \{\lambda \in \sigma_P(A_U) : \mathrm{Re}\lambda = 0\},
\end{aligned}$$

and assume $\Sigma_c \neq \emptyset$. We know that $\Sigma_c \cup \Sigma_u$ is a finite set.

Let

$$\begin{aligned}
C^s &= \bigoplus_{\lambda\in\Sigma_s} \mathcal{M}_\lambda(A_U), \\
C^u &= \bigoplus_{\lambda\in\Sigma_u} \mathcal{M}_\lambda(A_U), \\
C^c &= \bigoplus_{\lambda\in\Sigma_c} \mathcal{M}_\lambda(A_U).
\end{aligned}$$

C^s, C^u, and C^c are realified generalized eigenspaces associated with Σ_s, Σ_u, and Σ_c, respectively. Then C^u and C^c are finitely dimensional and

$$C = C^s \oplus C^u \oplus C^c.$$

Recall that C^s, C^u, and C^c are called the *stable, unstable*, and *center* subspaces of the C_0-semigroup $\{U(t)\}_{t\ge0}$.

We can now state the main result of this section.

THEOREM 6.2. *There exist open neighborhoods N_c of 0 in C^c, N_s of 0 in C^s, N^u of 0 in C^u, and a C^k-map $w_c : N_c \to C^s \oplus C^u$ such that*

(i) $w_c(0) = 0, Dw_c(0) = 0, w_c(N_c) \subset N_s + N_u$;

(ii) *for any $\phi \in V$, if there exists a continuous mapping $u^\phi : \mathbb{R} \to X$ such that $u_0^\phi = \phi$,*

$$u^\phi(t) = T(t-s)u^\phi(s) + \int_s^t T(t-\theta)[L(u_\theta^\phi) + F(u_\theta^\phi)]d\theta$$

for $t, s \in \mathbb{R}$ with $t \geq s$, and $u_t^\phi \in N_s + N_u + N_c$ for all $t \in \mathbb{R}$, then $u_t^\phi \in W_c$ for $t \in \mathbb{R}$, where

$$W_c = \{\phi_c + w_c(\phi_c) : \phi_c \in N_c\}.$$

Proof. Recall that $f : V \to C$ is C^k-smooth, $f(0) = 0$, $Df(0) = U(\omega)$, and

$$\begin{cases} C = C^s \oplus C^u \oplus C^c, \\ U(\omega)C^s \subset C^s,\ U(\omega)C^u \subset C^u,\ U(\omega)C^c \subset C^c, \\ \sigma(U(\omega)|_{C^s}) \text{ is a compact subset of } \{z \in \mathbb{C} : |z| < 1\}, \\ \sigma(U(\omega)|_{C^c}) \subset S^1_{\mathbb{C}}, \\ \sigma(U(\omega)|_{C^u}) \subset \{z \in \mathbb{C} : |z| > 1\}. \end{cases}$$

See Chapter IV.2 in Diekmann et al. [4].

By Theorem 5.4, there exist open neighborhoods N_c of 0 in C^c, N_s of 0 in C^s, N_u of 0 in C^u, and a C^k-map $w : N_c \to C^s \oplus C^u$ such that $w_c(0) = 0$, $Dw_c(0) = 0$, and $w_c(N_c) \subset N_s + N_u$. Moreover, for $W_c = \{\phi_c + w_c(\phi_c) : \phi_c \in N_c\}$, if there exists $(\phi^n)_{-\infty}^{\infty}$ such that $\phi^n = f(\phi^{n-1})$ and $\phi^n \in N_c + N_{su}$ for $n \in \mathbf{Z}$, then $\phi^0 \in W_c$.

Fix $\phi \in V$ such that condition (ii) of this theorem is satisfied. Then for any fixed $t \in \mathbb{R}$, $u_t^\phi \in N_s + N_u + N_c \subset V$, and if we let

$$\phi^n = u_{t+n\omega}^\phi, \qquad n \in \mathbf{Z},$$

then $\phi^{n+1} = f(\phi^n)$ for $n \in \mathbf{Z}$ and $\phi^n \in N_c + N_s + N_u$ for all $n \in \mathbf{Z}$. Therefore, the result in the last step implies that $\phi^0 = u_t^\phi \in W_c$. This completes the proof. ∎

Acknowledgment. The authors thank one of the referees, whose careful reading and valuable comments enabled them to improve the first version of this paper.

REFERENCES

[1] O. ARINO AND E. SANCHEZ, *Linear theory of abstract functional differential equations of retarded type,* J. Math. Anal. Appl., 191 (1995), pp. 547–571.

[2] S. BUSENBERG AND W. HUANG, *Stability and Hopf bifurcation for a population delay model with diffusion effects,* J. Differential Equations, 124 (1996), pp. 80–107.

[3] J. CARR, *Applications of Center Manifold Theory*, Springer-Verlag, New York, 1981.

[4] O. DIEKMANN, S. A. VAN GILS, S. M. VERDUYN LUNEL, AND H.-O. WALTHER, *Delay Equations. Functional, Complex, and Nonlinear Analysis*, Springer-Verlag, New York, 1995.

[5] T. FARIA, *Bifurcations aspects for some delayed population models with diffusion*, in Differential Equations with Applications to Biology, Fields Inst. Commun. 21, S. Ruan, G. Wolkowicz, and J. Wu, eds., AMS, Providence, RI, 1999, pp. 143–158.

[6] W. FITZGIBBON, *Semilinear functional differential equations in Banach spaces,* J. Differential Equations, 29 (1978), pp. 1–14.

[7] G. Greiner, *Compact and quasi-compact semigroups*, in One-Parameter Semigroups of Linear Operators, Lectures Notes in Math. 1184, R. Nagel, ed., Springer-Verlag, Berlin, New York, 1986, pp. 209–218.
[8] J. K. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, Berlin, 1977.
[9] W. Huang, *Studies in Differential Equations and Applications*, Ph.D. dissertation, Claremont Graduate School, University Microfilms International, 1990.
[10] T. Krisztin, H.-O. Walther, and J. Wu, *Shape, Smoothness and Invariant Stratification of an Attracting Set for Delayed Monotone Positive Feedback*, Fields Inst. Monogr. 11, AMS, Providence, RI, 1999.
[11] X. Lin, J. W.-H. So, and J. Wu, *Centre manifolds for partial differential equations with delays,* Proc. Roy. Soc. Edinburgh Sect. A, 122 (1992), pp. 237–254.
[12] M. C. Memory, *Stable and unstable manifolds for partial functional differential equations,* Nonlinear Anal., 16 (1991), pp. 131–142.
[13] S.-I. Nakagiri, *Structural properties of functional differential equations in Banach spaces,* Osaka J. Math., 25 (1988), pp. 353–398.
[14] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[15] K. Schumacher, *On the resolvent of linear nonautonomous partial functional differential equations,* J. Differential Equations, 59 (1985), pp. 355–387.
[16] J.-S. Shin and T. Naito, *Semi-Fredholm operators and periodic solutions for linear functional differential equations,* J. Differential Equations, 153 (1999), pp. 407–441.
[17] A. E. Taylor and D. C. Lay, *Introduction to Functional Analysis*, Wiley, New York, 1980.
[18] C. C. Travis and G. F. Webb, *Existence and stability for partial functional differential equations,* Trans. Amer. Math. Soc., 200 (1974), pp. 395–418.
[19] A. Vanderbauwhede and G. Iooss, *Center manifold theory in infinite dimensions*, in Dynamics Reported: Expositions in Dynamical Systems, Dynam. Report. Expositions Dynam. Systems (N.S.) 1, C. K. R. T. Jones, U. Kirchgraber, and H.-O. Walther, eds., Springer-Verlag, New York, 1992, pp. 125–163.
[20] A. Vanderbauwhede and S. A. van Gils, *Center manifolds and contractions on a scale of Banach spaces,* J. Funct. Anal., 71 (1987), pp. 209–224.
[21] J. Wu, *Theory and Applications of Partial Functional Differential Equations*, Springer-Verlag, New York, 1996.
[22] M. Yamamoto and S.-I. Nakagiri, *Identifiability of operators for evolution equations in Banach spaces with an application to transport equations,* J. Math. Anal. Appl., 186 (1994), p. 161–181.
[23] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, Vol. I, Springer-Verlag, New York, 1986.

SIAM J. MATH. ANAL.
Vol. 34, No. 1, pp. 204–227

UNIFORM PERSISTENCE, COEXISTENCE, AND EXTINCTION IN ALMOST PERIODIC/NONAUTONOMOUS COMPETITION DIFFUSION SYSTEMS*

GEORG HETZER† AND WENXIAN SHEN†

Dedicated to Professor Paul Waltman on the occasion of his 70th birthday

Abstract. A two species competition model with diffusion is considered. The parameters describing the growth, interaction, and self-limitation of the species are spatially inhomogeneous and temporally almost periodic. The boundary conditions are homogeneous and of Neumann or Dirichlet type. First, a convergence theorem is derived in the single species case. Roughly speaking, it states that one of the following alternatives will occur: either every positive solution converges to a unique strictly positive almost periodic solution, every positive solution converges to the trivial solution, or every positive solution is neither bounded away from the trivial solution nor converges to it. Then appropriate conditions for uniform persistence of both species as well as for extinction of one of the species are established. Moreover, it is shown that uniform persistence implies coexistence in the sense that there is a strictly positive solution whose hull is almost automorphic. The above results generalize earlier work in the time independent and time periodic cases for both single species population models and two species competition models. The approach developed in this paper for dealing with almost periodic equations can be applied to more general nonautonomous equations, as we will indicate by briefly discussing applications where merely time recurrence is supposed.

Key words. almost periodicity, almost automorphy, competition, uniform persistence, coexistence, extinction

AMS subject classifications. 35B15, 35B40, 35K57, 92D25

PII. S0036141001390695

1. Introduction.

A central issue in population dynamics is the long-term development of populations, and one finds terms such as uniform persistence (sometimes also called permanence), coexistence, and extinction describing important special types of asymptotic behavior of the solutions of associated model equations. The time independent and time periodic cases have found great interest in the past. Our goal, however, is to investigate possible scenarios for two species populations in the case where the model equations depend on time nonperiodically. The reaction-diffusion system under consideration is given by

$$\begin{cases} u_t = k_1\Delta u + u(a_1(t,x) - b_1(t,x)u - c_1(t,x)v), & x \in \Omega, \\ v_t = k_2\Delta v + v(a_2(t,x) - b_2(t,x)u - c_2(t,x)v), & x \in \Omega, \\ Bu = Bv = 0, & x \in \partial\Omega, \end{cases} \tag{1.1}$$

where k_1, k_2 are positive constants, a_i, b_i, c_i $(i = 1, 2)$ are smooth functions, $\Omega \subset \mathbb{R}^n$ is a smooth bounded region, and $Bu = \frac{\partial u}{\partial n}$, $Bv = \frac{\partial v}{\partial n}$ or $Bu = u$, $Bv = v$.

System (1.1) models the competition between two species. It is sometimes called the Lotka–Volterra competition model. In the context of ecology, k_1, k_2 are the dispersal rates, a_1, a_2 represent growth rates, b_1, c_2 denote self-limitation rates, and c_1, b_2 are the interaction rates. Dirichlet boundary conditions ($u = v = 0$, $x \in \partial\Omega$)

*Received by the editors June 10, 2001; accepted for publication (in revised form) March 26, 2002; published electronically September 5, 2002.

http://www.siam.org/journals/sima/34-1/39069.html

†Department of Mathematics, Auburn University, Auburn, AL 36849 (hetzege@mail.auburn.edu, ws@math.auburn.edu). The second author was partially supported by NSF grant DMS-9704245.

describe a "lethal crossing" boundary, and Neumann boundary conditions ($\frac{\partial u}{\partial n} = \frac{\partial v}{\partial n} = 0$) exclude migration across the boundary.

Ecologically, one is only interested in positive solutions of system (1.1), and it is the objective of this paper to study uniform persistence, coexistence, and extinction for (1.1). Roughly speaking, *uniform persistence* means that there are strictly positive functions $u_*(\cdot)$, $u^*(\cdot)$, $v_*(\cdot)$, and $v^*(\cdot)$ such that, for every positive solution $(u(t,x), v(t,x))$ of (1.1), $u_*(x) \le u(t,x) \le u^*(x)$, and $v_*(x) \le v(t,x) \le v^*(x)$ for $x \in \Omega$ and large t, *coexistence* refers to the existence of certain distinct strictly positive solutions, and *extinction* indicates that at least one of the species eventually dies out. These issues have been studied widely for time independent or time periodic equations under Neumann boundary conditions; cf. [3], [4], [6], [9], [13], [16], [20], [25], [26], [27], [30], [31], [42], [46], [47], [48], etc. In [8], [10], etc., one also finds results in the case of Dirichlet boundary conditions. It turns out that, thanks to the Poincaré map, time periodic equations resemble time independent ones in many aspects.

In nature, populations evolve influenced by external effects which are roughly, but not exactly periodic, or under environmental forcing which exhibits different, noncommensurate periods. This sort of time dependence can arise from the interplay of short-term weather cycles and seasonal climate variations, or from the superposition of daily and annually periodic phenomena, and so on. Growth processes, for example, depend on the length of days and nights which varies during the year. Models with such time dependence are characterized more appropriately by quasi-periodic or almost periodic equations or even by certain nonautonomous equations rather than by periodic ones. Additionally, populations are affected by a wide variety of irregularly occurring phenomena which lead to stochastic or random equations. Both types of equations, time nonperiodic deterministic (e.g., quasi-periodic and almost periodic) equations and time stochastic ones are therefore worth studying. In this paper, we will focus on the first type of equations, in particular, time quasi-periodic, almost periodic, and certain general nonautonomous equations. These equations have found much attention (see [1], [2], [14], [15], [22], [37], etc.). However, in contrast to the time independent and periodic cases, many fundamental questions about general nonautonomous, even about quasi-periodic or almost periodic cases remain open. For example, in the time periodic case, uniform persistence implies the coexistence in the sense that there is a strictly positive periodic solution having the same period as the period of the functions arising in (1.1). However, it is not known yet whether, in the general time quasi-periodic (almost periodic) case, uniform persistence implies the existence of a strictly positive quasi-periodic (almost periodic) solution. If not, which kind of distinct positive solution may result from uniform persistence? Note that quasi periodicity is a special kind of almost periodicity. It is known (see [38]) that in a certain sense quasi-periodic time dependence does not result in "better" dynamics than general almost periodic dependence. Motivated by these facts, we will mainly focus in this paper on (1.1) with general almost periodic time dependence. Typically, studies about periodic equations are carried out in terms of the Poincaré map. A unified framework to study an almost periodic equation is the so-called skew-product (semi)flow generated by the equation (see [35], [36], [38], etc.). We will carry out our study for almost periodic competition models in the skew-product setting, but note that the approach developed in this paper for dealing with almost periodic equations can be applied to more general time dependent settings. Moreover, methodologically this framework can also be considered a precursor for a study of associated stochastic

problems, even if the concept of random dynamical systems would prove to be inadequate due to technical obstacles from stochastic partial differential equations (cocycle property).

From now on, we assume that a_i, b_i, c_i $(i = 1, 2)$ in (1.1) are smooth functions (i.e., continuous in x on $\bar{\Omega}$ and Hölder continuous in t) and that they are uniformly almost periodic in t (see section 2.1 for definition) unless otherwise specified. In such a setting, system (1.1) has been considered under Neumann boundary conditions in [1], [2], [14], etc. In [22], we studied the long-time behavior of positive solutions of (1.1) when a_i, b_i, c_i are spatially homogeneous and Neumann boundary conditions are prescribed. We took advantage of findings for the associated system of ODEs,

$$\begin{cases} \dot{u} = u(a_1(t) - b_1(t)u - c_1(t)v) \\ \dot{v} = v(a_2(t) - b_2(t)u - c_2(t)v), \end{cases} \tag{1.2}$$

by investigating the relation between (1.1) and (1.2) and utilized convergence results from [37] for the single species population model

$$\begin{cases} u_t = k\Delta u + u(a(t,x) - b(t,x)u), & x \in \Omega, \\ Bu = 0, \quad x \in \partial\Omega, \end{cases} \tag{1.3}$$

assuming that $Bu = \frac{\partial u}{\partial n}$, k is a positive constant, a and b are smooth functions which are uniformly almost periodic in t, and Ω has the same meaning as for (1.1).

Clearly, a new approach is required when dealing with (1.1) in the general case. As a first step, we develop a general method for studying the convergence of positive solutions of the single species population model (1.3) when $Bu = u$ or $Bu = \frac{\partial u}{\partial n}$. A solution $u(t,x)$ of (1.3) is said to be *positive* if $u(t,x) \geq 0$ $(u \not\equiv 0)$ for $x \in \Omega$ and $t \geq 0$, and *strictly positive* if in the Neumann case $(Bu = \frac{\partial u}{\partial n})$ one has $u(x,t) > 0$ for $x \in \bar{\Omega}$ and $t \geq 0$, and if in the Dirichlet case $(Bu = u)$ one has $u(x,t) > 0$ for $x \in \Omega$ and $t \geq 0$, $\frac{\partial u}{\partial n} < 0$ for $x \in \partial\Omega$ and $t \geq 0$. A solution $u(t,x)$ of (1.3) is said to *converge* to $u^*(t,x)$ if $u(t,\cdot) - u^*(t,\cdot) \to 0$ in an appropriate function space norm as $t \to \infty$. Denote by $\mathcal{M}(\cdot)$ the frequency module of an almost periodic (almost automorphic) function (see section 2.1 for a definition). We prove the following result.

THEOREM A (Corollary 3.4). *Consider* (1.3) *and assume* $b(t,x) \geq \delta$ *for some* $\delta > 0$. *One and only one of the following alternatives occurs.*

(1) *Every positive solution converges to a unique strictly positive almost periodic solution* $u^*(t,x)$ *with* $\mathcal{M}(u^*) \subset \mathcal{M}(a,b)$.

(2) *Every positive solution converges to the trivial solution* $u = 0$.

(3) *Every positive solution is neither bounded away from the trivial solution nor converges to it.*

Notice that if a, b in (1.3) are actually periodic in t, alternative (3) of Theorem A cannot occur in view of the generic convergence of monotone dynamical systems ([24], [29], [33], [40], etc.). Theorem A extends the convergence results proved for the periodic case in [18], [19], [21], as well as the convergence results for (1.3) which were obtained in [37] for the Neumann case $Bu = \frac{\partial u}{\partial n}$. It should be pointed out that Takáč [43] has obtained some results which partially cover Theorem A, by using the so-called part metric.

Next, we consider positive solutions of (1.1). A solution $(u(t,x), v(t,x))$ of (1.1) is said to be *positive* if $u(t,x) \geq 0$, $v(t,x) \geq 0$ $(u(t,x) \not\equiv 0$, $v(t,x) \not\equiv 0)$ for $x \in \Omega$ and $t \geq 0$, and *strictly positive* if one has in the Neumann case that $u(t,x) > 0$, $v(t,x) > 0$ for $x \in \bar{\Omega}$ and $t \geq 0$, and if one has in the Dirichlet case that $u(t,x) > 0$, $v(t,x) > 0$

for $x \in \Omega$ and $t \geq 0$, $\frac{\partial u}{\partial n} < 0$, $\frac{\partial v}{\partial n} < 0$ for $x \in \partial\Omega$ and $t \geq 0$. A solution $(u(t,x), v(t,x))$ of (1.1) is said to *converge* to $(u^*(t,x), v^*(t,x))$ if, in a certain function space norm, $u(t,x) - u^*(t,x) \to 0$ and $v(t,x) - v^*(t,x) \to 0$ as $t \to \infty$. Let

$$a_{iL(M)} = \inf(\sup)_{t\in\mathbb{R}, x\in\bar{\Omega}} a_i(t,x), \tag{1.4}_1$$

$$b_{iL(M)} = \inf(\sup)_{t\in\mathbb{R}, x\in\bar{\Omega}} b_i(t,x), \tag{1.4}_2$$

$$c_{iL(M)} = \inf(\sup)_{t\in\mathbb{R}, x\in\bar{\Omega}} c_i(t,x). \tag{1.4}_3$$

Assume that a_{iL}, b_{iL}, $c_{iL} > 0$, and $u \equiv 0$ is an unstable solution of

$$\begin{cases} u_t = k_1\Delta u + u(a_1 - b_1 u), & x \in \Omega, \\ Bu = 0, & x \in \partial\Omega \end{cases} \tag{1.5}$$

and $v \equiv 0$ is an unstable solution of

$$\begin{cases} v_t = k_2\Delta v + v(a_2 - c_2 v), & x \in \Omega, \\ Bv = 0, & x \in \partial\Omega. \end{cases} \tag{1.6}$$

Then, by Theorem A(1), (1.5), (1.6) has a globally stable strictly positive almost periodic solution $u^*(t,x)$ $(v^*(t,x))$ with $\mathcal{M}(u^*) \subset \mathcal{M}(a_1, b_1)$ $(\mathcal{M}(v^*) \subset \mathcal{M}(a_2, c_2))$. We show the following results.

THEOREM B (Theorem 5.1). *Consider* (1.1). *Assume that* $Bu = \frac{\partial u}{\partial n}$.

(1) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$ *and* $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, *then uniform persistence occurs. Moreover, there is a strictly positive solution* $(u_*(t,x), v_*(t,x))$ *whose hull is almost automorphic, and, for each* $(\tilde{u}_*(\cdot,\cdot), \tilde{v}_*(\cdot,\cdot)) \in H(u_*, v_*)$ *with* $\tilde{u}_*(t,x)$, $\tilde{v}_*(t,x)$ *being almost automorphic in* t, $\mathcal{M}(\tilde{u}_*, \tilde{v}_*) \subset \mathcal{M}(a_1, b_1, c_1, a_2, b_2, c_2)$.

(2) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$ *and* $a_{2M} \leq \frac{a_{1L}b_{2L}}{b_{1M}}$, *then every positive solution converges to* $(u^*(t,x), 0)$.

(3) *If* $a_{1M} \leq \frac{c_{1L}a_{2L}}{c_{2M}}$ *and* $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, *then every positive solution converges to* $(0, v^*(t,x))$.

(4) *If* $a_1 = a_2$, $b_1 = b_2 = c_1 = c_2$, *and also* $k_1 = k_2$ *in the case where* a_i, b_i, c_i $(i = 1, 2)$ *are not spatially homogeneous, then there is a stable continuous family of strictly positive almost periodic solutions connecting* $(u^*(t,x), 0)$ *and* $(0, v^*(t,x))$.

THEOREM C. (Theorem 5.2). *Consider* (1.1). *Assume that* $Bu = u$.

(1) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$, $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, $k_1 = k_2$, *and* $a_1 = a_2$ *(constant), then uniform persistence occurs. Moreover, there is a strictly positive solution* $(u_*(t,x), v_*(t,x))$ *whose hull is almost automorphic, and one has* $\mathcal{M}(\tilde{u}_*, \tilde{v}_*) \subset \mathcal{M}(a_1, b_1, c_1, a_2, b_2, c_2)$ *for each* $(\tilde{u}_*(\cdot,\cdot), \tilde{v}_*(\cdot,\cdot)) \in H(u_*, v_*)$ *for which* $\tilde{u}_*(\cdot,\cdot)$, $\tilde{v}_*(\cdot,\cdot)$ *are almost automorphic in* t.

(2) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$, $a_{2M} \leq \frac{a_{1L}b_{2L}}{b_{1M}}$, $k_2 \geq k_1$, *and* $a_{1L} \geq a_2$, *then every positive solution converges to* $(u^*(t,x), 0)$.

(3) *If* $a_{1M} \leq \frac{c_{1L}a_{2L}}{c_{2M}}$, $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, $k_1 \geq k_2$, *and* $a_{2L} \geq a_1$, *then every positive solution converges to* $(0, v^*(t,x))$.

(4) *If* $k_1 = k_2$, $a_1 = a_2$, *and* $b_1 = b_2 = c_1 = c_2$, *then there is a stable continuous family of strictly positive almost periodic solutions connecting* $(u^*(t,x), 0)$ *and* $(0, v^*(t,x))$.

As pointed out before, in the periodic case, uniform persistence implies coexistence in the sense that there is a strictly positive periodic solution with the same period as the period of the functions arising in (1.1). By Theorems B and C, in the almost periodic case, uniform persistence implies the coexistence in the sense that there is a

strictly positive solution whose hull is almost automorphic with the same frequency module as that of the functions arising in (1.1). Note that such a solution is periodic if these functions are actually periodic and all have the same period. Therefore, on the one hand, Theorems B and C shed light on uniform persistence, coexistence, and extinction in the case of almost periodic competition models, and, on the other hand, they generalize most existing results for corresponding time independent or time periodic competition models.

Also, as previously mentioned, the approach which we develop here for dealing with (1.1) and (1.3) in the case of almost periodic time dependence extends to more general situations. For example, if we assume only recurrent time dependence (see section 2.1 for a definition) in the context of (1.1) and (1.3), we obtain the following results.

THEOREM D (Theorem 6.1). *Consider* (1.3) *with a, b being recurrent in t and $b \geq \delta$ for some $\delta > 0$. One and only one of the following alternatives occurs.*

(1) *Every positive solution converges to a unique strictly positive recurrent solution $u^*(t,x)$ whose hull is a* 1*-cover of the hull of (a,b).*

(2) *Every positive solution converges to the trivial solution $u = 0$.*

(3) *Every positive solution is neither bounded away from the trivial solution nor converges to it.*

THEOREM E (Theorem 6.2). Consider (1.1) with a_i, b_i, c_i $(i = 1, 2)$ being recurrent in t. Assume that $Bu = \frac{\partial u}{\partial n} = 0$.

(1) If $a_{1L} > \frac{c_{1M} a_{2M}}{c_{2L}}$ and $a_{2L} > \frac{a_{1M} b_{2M}}{b_{1L}}$, then uniform persistence occurs. Moreover, there is a strictly positive recurrent solution $(u_*(t,x), v_*(t,x))$ whose hull is an almost 1-cover of the hull of $(a_1, b_1, c_1, a_2, b_2, c_2)$.

(2) If $a_{1L} > \frac{c_{1M} a_{2M}}{c_{2L}}$ and $a_{2M} \leq \frac{a_{1L} b_{2L}}{b_{1M}}$, then every positive solution converges to $(u^*(t,x), 0)$, where $u^*(t,x)$ is the unique strictly positive recurrent solution of (1.5) guaranteed by Theorem D(1).

(3) If $a_{1M} \leq \frac{c_{1L} a_{2L}}{c_{2M}}$ and $a_{2L} > \frac{a_{1M} b_{2M}}{b_{1L}}$, then every positive solution converges to $(0, v^*(t,x))$, where $v^*(t,x)$ is the unique strictly positive recurrent solution of (1.6) guaranteed by Theorem D(1).

THEOREM F (Theorem 6.3). *Consider* (1.1) *with a_i, b_i, c_i $(i = 1, 2)$ being recurrent in t. Assume that $Bu = u$.*

(1) *If $a_{1L} > \frac{c_{1M} a_{2M}}{c_{2L}}$, $a_{2L} > \frac{a_{1M} b_{2M}}{b_{1L}}$, $k_1 = k_2$, and $a_1 = a_2$ (constant), then uniform persistence occurs. Moreover, there is a strictly positive recurrent solution $(u_*(t,x), v_*(t,x))$ whose hull is an almost* 1*-cover of the hull of $(a_1, b_1, c_1, a_2, b_2, c_2)$.*

(2) *If $a_{1L} > \frac{c_{1M} a_{2M}}{c_{2L}}$, $a_{2M} \leq \frac{a_{1L} b_{2L}}{b_{1M}}$, $k_2 \geq k_1$, and $a_{1L} \geq a_2$, then every positive solution converges to $(u^*(t,x), 0)$, where $u^*(t,x)$ is the unique strictly positive recurrent solution of* (1.5) *guaranteed by Theorem* D(1).

(3) *If $a_{1M} \leq \frac{c_{1L} a_{2L}}{c_{2M}}$, $a_{2L} > \frac{a_{1M} b_{2M}}{b_{1L}}$, $k_1 \geq k_2$, and $a_{2L} \geq a_1$, then every positive solution converges to $(0, v^*(t,x))$, where $v^*(t,x)$ is the unique strictly positive recurrent solution of* (1.6) *guaranteed by Theorem* D(1).

Let us end the discussion of our results by the following remarks. First, Theorems D, E, and F apply to both the almost periodic and the almost automorphic cases, and, when applied to the almost periodic case, Theorems A, B, and C, respectively, are recovered. In order to be more precise, if a and b are actually uniformly almost periodic (almost automorphic) in t (see section 2.1 for a definition), $u^*(t,x)$ in Theorem D(1) is also almost periodic (almost automorphic) and $\mathcal{M}(u^*) \subset \mathcal{M}(a,b)$. In the case where a_i, b_i, and c_i $(i = 1, 2)$ are uniformly almost periodic (almost automorphic) in t, $u^*(t,x)$ and $v^*(t,x)$ in Theorems E, F(2) and Theorems E, F(3) are almost periodic

(almost automorphic), but $(u_*(t,x), v_*(t,x))$ in Theorems E, F(1) may be neither almost periodic nor almost automorphic; nevertheless, their hull is almost automorphic (see section 2.1 for a definition). Second, we remark that the approach developed in the current paper applies to more general right-hand sides of one species population and two species competition models (see [22] for the spatially homogeneous case). It also has certain implications on the study of stochastic population models (see [23] for one species population models). Finally, we note that the dynamical system framework employed in this paper can also be applied to (two species) competition in stirred and unstirred chemostats, though due to the lack of monotonicity many techniques developed in this paper do not extend to such models, and new techniques need to be introduced. A characteristic feature of such models is a nutrition in- and out-flow explicitly represented by an extra differential equation. Waltman has made numerous contributions to this area, but we refer here only to the monograph [41].

The paper is organized as follows. In section 2, we present some preliminary lemmas. Section 3 is devoted to the investigation of a single almost periodic population model, and Theorem A will be proved in this section. We establish some basic properties of the almost periodic two species competition model in section 4 and establish Theorems B and C in section 5. In section 6, we describe some results which can be obtained for (1.1) and (1.3) in the case of general recurrent time dependence by employing the approach which we have developed in this paper in the context of almost periodic time dependence.

2. Preliminary lemmas. In this section, we present some results about frequency module containment of almost periodic (almost automorphic) functions, ω-limit sets of skew-product semiflows, and spectra of linear scalar parabolic equations for use in later sections.

2.1. Almost periodic, almost automorphic, and recurrent functions. Let $E \subset \mathbb{R}^n$ and $f \in C(\mathbb{R}\times E, \mathbb{R}^m)$ be uniformly almost periodic (almost automorphic) in t. Recall that $f \in C(\mathbb{R}\times E, \mathbb{R}^m)$ is *uniformly almost periodic (almost automorphic) in* t if f is uniformly continuous on $\mathbb{R} \times E_0$ for any bounded subset $E_0 \subset E$ and is almost periodic (almost automorphic) in t for each $x \in E$, and $f(t,x)$ is *almost periodic (almost automorphic) in* t for given $x \in E$ if, for all sequences $\{\alpha_n'\}, \{\beta_n'\} \subset \mathbb{R}$ ($\{\alpha_n'\} \subset \mathbb{R}$), there are subsequences $\{\alpha_n\} \subset \{\alpha_n'\}$, $\{\beta_n\} \subset \{\beta_n'\}$ ($\{\alpha_n\} \subset \{\alpha_n'\}$) such that $\lim_{k\to\infty}\lim_{n\to\infty} f(t+\alpha_n+\beta_k, x) = \lim_{n\to\infty} f(t+\alpha_n+\beta_n, x)$ ($\lim_{k\to\infty}\lim_{n\to\infty} f(t+\alpha_n-\alpha_k, x) = f(t,x)$) pointwise for $t \in \mathbb{R}$ (see [11], [45] for details). The so-called frequency module of f is defined as follows. Let

$$f(t,x) \sim \sum_{\lambda\in\mathbb{R}} a_\lambda(x) e^{i\lambda t} \tag{2.1}$$

be a Fourier series of f (see [44], [45] for the definition and existence of Fourier series). Then $\mathcal{S}(f) = \{\lambda : a_\lambda(x) \not\equiv 0\}$ is called the *Fourier spectrum of* f *associated with the Fourier series* (2.3), and $\mathcal{M}(f) =$ the smallest additive subgroup of $\mathbb{R}$ containing $\mathcal{S}(f)$ is called the *frequency module of* f. We note that $\mathcal{S}(f)$ may depend on the chosen Fourier series (2.1), but $\mathcal{M}(f)$ is independent of that series (see [45]).

Let $E \subset \mathbb{R}^n$ and $f \in C(\mathbb{R}\times E, \mathbb{R}^n)$ be uniformly continuous on $\mathbb{R} \times E_0$ for each bounded subset $E_0 \subset E$. The hull $H(f)$ of f is defined to be $H(f) = cl\{f\cdot\tau | \tau \in \mathbb{R}\}$, where $f\cdot\tau(t,x) = f(t+\tau, x)$, and the closure is taken under the compact open topology. Denote $(H(f), \mathbb{R})$ as the translation flow $(g,t) = g\cdot t$ for each $g \in H(h)$ and $t \in \mathbb{R}$. f is said to have an *almost periodic* (*almost automorphic*) *hull* if $(H(f), \mathbb{R})$

is minimal and there is $g \in H(f)$ such that g is uniformly almost periodic (almost automorphic) in t. f is said to be *recurrent* if $(H(f), \mathbb{R})$ is minimal. Notice that if f is almost periodic (almost automorphic), then $(H(f), \mathbb{R})$ is minimal, and hence f is recurrent.

LEMMA 2.1. *Let $f(t,x)$ and $g(t,y)$ ($f \in C(\mathbb{R} \times E, \mathbb{R}^m)$, $g \in C(\mathbb{R} \times E, \mathbb{R}^k)$) be two uniformly almost automorphic functions in t. Then $\mathcal{M}(g) \subset \mathcal{M}(f)$ iff for each sequence $\{\alpha_n\} \subset \mathbb{R}$, if $\lim_{n\to\infty} f(t+\alpha_n, x) = f(t,x)$ uniformly for t and x in bounded sets, then $\lim_{n\to\infty} g(t+\alpha_n, y) = g(t,y)$ uniformly for t and y in bounded sets.*

Proof. See [45]. □

2.2. Skew-product semiflows. Let $(Y, \mathbb{R})$ be a minimal flow with Y being a compact metric space and $y \cdot t \equiv (y,t)$ for any $y \in Y$ and $t \in \mathbb{R}$. Let Z be a complete metric space and $\leq$ be a partial ordering on Z satisfying that if $z_n^1 \leq z_n^2$ ($n \in \mathbb{N}$), then $\lim_{n\to\infty} z_n^1 \leq \lim_{n\to\infty} z_n^2$ provided that both limits exist. Let $P : Z \times Y \to Y$ be the natural projection and $\pi_t : Z \times Y \to Z \times Y$ be a skew-product semiflow, that is, a semiflow of the form

$$\pi_t(z,y) = (\Psi(t;z,y), y \cdot t)$$

for all $(z,y) \in Z \times Y$ and $t \geq 0$. Given $(z_0, y_0) \in Z \times Y$, if $\{\pi_t(z_0,y_0) | t \geq t_0\}$ is relatively compact for a $t_0 > 0$, then the set

$$\omega(y_0, z_0) = \bigcap_{\tau \geq t_0} cl\{\pi_{t+\tau}(z_0,y_0) | t \geq 0\}$$

is called the *ω-limit set of $\pi_t(z_0,y_0)$*. We say π_t is *partially monotone with respect to* $\leq$ if, for any $y \in Y$ and any z_1, $z_2 \in Z$ with $z_1 \leq z_2$, $\Psi(t;z_1,y) \leq \Psi(t;z_2,y)$ for $t \geq 0$.

LEMMA 2.2. *Assume that, for each $(z,y) \in Z \times Y$, $\pi_t(z,y)$ has at most one backward extension and that $(z_0,y_0) \in Z \times Y$ is such that $\{\pi_t(z_0,y_0) | t \geq t_0\}$ is relatively compact for some $t_0 > 0$.*

(1) π_t has a flow extension on the ω-limit set $\omega(z_0,y_0)$.

(2) There is a residual invariant subset $Y_0 \subset Y$ such that, for all $y^* \in Y_0$, $y \in Y$, and $\{t_n\} \subset \mathbb{R}$ with $y \cdot t_n \to y^*$ and for all $(z^*, y^*) \in P^{-1}(y^*) \cap \omega(z_0,y_0)$, there is a sequence $\{(z_n, y)\} \subset P^{-1}(y) \cap \omega(z_0,y_0)$ such that $\pi_{t_n}(z_n, y) \to (z^*, y^*)$.

(3) If π_t is partially monotone with respect to $\leq$, and $z_0 \leq z$ ($z \leq z_0$) for every $(z, y_0) \in \omega(z_0,y_0)$, then $\omega(z_0,y_0)$ is an almost 1-cover of Y; that is, $P^{-1}(y) \cap \omega(z_0,y_0)$ is a singleton for residually many $y \in Y$.

Proof. (1) The proof follows from [38].

(2) See [39] or [45].

(3) Without loss of generality, assume that $z_0 \leq z$ for every $(z,y_0) \in \omega(z_0,y_0)$. Let $Y_0 \subset Y$ be the set guaranteed by (2). For $y^* \in Y_0$ and (z_1,y^*), $(z_2,y^*) \in \omega(z_0,y_0)$, let $t_n \to \infty$ be such that $\pi_{t_n}(z_0,y_0) \to (z_1,y^*)$. By (2), one finds $(z^n, y_0) \in \omega(z_0,y_0)$ ($n \in \mathbb{N}$) such that $\pi_{t_n}(z^n,y_0) \to (z_2,y^*)$. Since $z_0 \leq z^n$, one has $\pi_{t_n}(z_0,y_0) \leq \pi_{t_n}(z^n,y_0)$ for $n \geq 1$, thus $z_1 \leq z_2$. Similarly, we can prove that $z_2 \leq z_1$. Hence $z_1 = z_2$ and $\omega(z_0,y_0)$ is an almost 1-cover of Y. □

2.3. Spectrum of linear scalar parabolic equations. Given a smooth bounded region $\Omega \subset \mathbb{R}^n$, let $X \subset L^p(\Omega)$ ($p > n$) be a fractional power space of $-\Delta : \mathcal{D} \to L^p(\Omega)$ satisfying $X \hookrightarrow C^1(\bar{\Omega})$, where $\mathcal{D} = \{u \in H^{2,p}(\Omega) | Bu = 0 \text{ for } x \in \partial\Omega\}$ and $Bu = \frac{\partial u}{\partial n}$ or u [17]. Then $\mathrm{Int}X_+ \neq \emptyset$. This is because, in the case that $Bu = \frac{\partial u}{\partial n}$, the set $\{u \in X | u(x) > 0 \text{ for } x \in \bar{\Omega}\} \subset \mathrm{Int}X_+$ is not empty, and, in the case

that $Bu = u$, the set $\{u \in X | u(x) > 0$ for $x \in \Omega$ and $\frac{\partial u}{\partial n} < 0$ for $x \in \partial\Omega\} \subset \mathrm{Int}X_+$ is not empty. Consequently, X_+ defines a strong ordering on X as follows:

$$u_1 \leq u_1 \quad \text{if} \quad u_1(x) \leq u_2(x) \quad \text{for} \quad \text{all} \quad x \in \Omega, \tag{2.2}$$

$$u_1 < u_2 \quad \text{if} \quad u_1 \leq u_2 \quad \text{but} \quad u_1 \neq u_2, \tag{2.3}$$

$$u_1 \ll u_2 \quad \text{if} \quad u_2 - u_1 \in \mathrm{Int}X_+. \tag{2.4}$$

Let $(Y, \mathbb{R})$ be a compact flow and $y \cdot t \equiv (y, t)$. Consider

$$\begin{cases} u_t = k\Delta u + A(y \cdot t, x)u, & x \in \Omega, \\ Bu = 0, & x \in \Omega, \end{cases} \tag{2.5}_y$$

where u is a scalar function, k is a positive constant, $y \in Y$, $A : Y \times \bar{\Omega} \to \mathbb{R}$ is continuous, and $A(y \cdot t, x)$ is Hölder continuous in t uniformly with respect to $y \in Y$ and $x \in \bar{\Omega}$. Then $(2.5)_y$ generates a skew-product semiflow [17], $\Pi_t : X \times Y \to X \times Y$,

$$\Pi_t(u, y) = (\Phi(t; u, y), y \cdot t), \tag{2.6}$$

where $\Phi(t; u, y)$ is the solution of $(2.5)_y$ with $\Phi(0; u, y) = u$. Thanks to the maximum principle for parabolic equations [12], [34], Π_t in (2.6) is strongly monotone in the sense that $\Phi(t; u, y) \gg 0$ for any $t > 0$, $y \in Y$, and $u \in X_+$ with $u \not\equiv 0$. The so-called Sacker–Sell spectrum and upper Lyapunov exponent of (2.6) or $(2.5)_y$ are defined as follows.

For a given $\sigma \in \mathbb{R}$, define $\Pi_t^\sigma : X \times Y \to X \times Y$, $t \geq 0$,

$$\Pi_t^\sigma(u, y) = (\Phi_\sigma(t; u, y), y \cdot t), \tag{2.7}_\sigma$$

where $\Phi_\sigma(t; u, y) = e^{-\sigma t}\Phi(t; u, y)$. Then the set

$$\Sigma(k, Y) = \{\sigma \in \mathbb{R} \mid (2.7)_\sigma \text{ admits no exponential dichotomy}\}$$

is called the *Sacker–Sell (dynamical) spectrum* and the number $\lambda(k, Y) = \sup_{y \in Y} \lambda(k, y)$ is called the *upper Lyapunov exponent* of (2.6), where $\lambda(k, y) = \limsup_{t \to \infty} \frac{\ln \|\Phi(t; \cdot, y)\|}{t}$.

Notice that, for any smooth function $h(t, x)$ which is uniformly almost periodic (almost automorphic, recurrent) in t, the almost periodic (almost automorphic, recurrent) scalar parabolic equation

$$\begin{cases} u_t = k\Delta u + h(t, x)u, & x \in \Omega, \\ Bu = 0, & x \in \partial\Omega, \end{cases} \tag{2.8}$$

can be built into $(2.5)_y$ by letting $Y = H(h)$ and $A : Y \times \bar{\Omega} \to \mathbb{R}$, $A(g, x) = g(0, x)$ for any $g \in Y$ and $x \in \bar{\Omega}$. Moreover, the dynamics of (2.8) is then reflected by that of the skew-product semiflow generated $(2.5)_y$ with $y \in Y = H(h)$. If $h(t, x) = h(x)$ is independent of t, then $\lambda(k, h)$ is the largest eigenvalue of the eigenvalue problem,

$$\begin{cases} k\Delta u + h(x)u = \lambda u, & x \in \Omega, \\ Bu = 0, & x \in \partial\Omega. \end{cases} \tag{2.9}$$

LEMMA 2.3. (1) *Suppose that $h_1(t, x)$ and $h_2(t, x)$ are two smooth functions which are uniformly almost periodic (almost automorphic, recurrent) in t. If $h_1(t, x) \leq h_2(t, x)$, then $\lambda(k, H(h_1)) \leq \lambda(k, H(h_2))$.*

(2) *If h is uniformly almost periodic in t, then $\lambda(k,g)$ is independent of $g \in H(h)$ (consequently, $\lambda(k,H(h)) = \lambda(k,h)$).*

(3) *If h is uniformly almost periodic in t, then $\lambda(k,h) \geq \lambda(k,\bar{h})$, where $\bar{h}(x) = \lim_{t\to\infty} \frac{1}{t}\int_0^t h(s,x)ds$.*

(4) *If h is uniformly almost periodic in t, then there is $\sigma_1 < \lambda(k,h)$ such that $\Sigma(k,H(h)) = \Sigma_1 \cup \{\lambda(k,h)\}$ and $\lambda \leq \sigma_1$ for any $\lambda \in \Sigma_1$.*

(5) *Suppose that $0 < k_1 \leq k_2$ are two constants and $h_1(x) \geq h_2(x)$ are two smooth functions. Then $\lambda(k_1,h_1) \geq \lambda(k_2,h_2)$, and the strict inequality holds if $h_1 \not\equiv h_2$.*

Proof. (1), (2), (3), and (4) follow from the arguments in [28] and the exponential separation theory in [32]. (5) follows from the fact that

$$\lambda(k_i,h_i) = \sup_{u\in X, \|u\|_2=1}\left(-k_i\int_\Omega |\bigtriangledown u(x)|^2dx + \int_\Omega h_i(x)u^2(x)dx\right)$$

$(i = 1,2)$, where $\|u\|_2 = (\int_\Omega u^2(x)dx)^{1/2}$. □

3. Convergence for almost periodic single species population models. In this section, we consider the convergence in the following single species population model:

$$\begin{cases} u_t = k\Delta u + u(a(t,x) - b(t,x)u), & x \in \Omega, \\ Bu = 0, & x \in \partial\Omega, \end{cases} \tag{3.1}$$

where k is a positive constant, $a(t,x)$, $b(t,x)$ $(b(t,x) \geq \delta > 0)$ are smooth functions and are uniformly almost periodic in t, $\Omega \subset \mathbb{R}^n$ is a smooth bounded region, and $Bu = \frac{\partial u}{\partial n}$ or u.

Let $X \hookrightarrow C^1(\bar{\Omega})$ be as in section 2.3, and let $\|\cdot\|$ be the norm of X. Let $P : X\times H(a,b) \to H(a,b)$ be the natural projection and $\Pi_t : X\times H(a,b) \to X\times H(a,b)$ be the (local) skew-product semiflow generated by (3.1); that is,

$$\Pi_t(u_0,a,b) = (u(t,\cdot;u_0,c,d), c\cdot t, d\cdot t), \tag{3.2}$$

where $u(t,\cdot;u_0,c,d)$ is the solution of

$$\begin{cases} u_t = k\Delta u + u(c(t,x) - d(t,x)u), & x \in \Omega, \\ Bu = 0, & x \in \partial\Omega \end{cases} \tag{3.3}_{c,d}$$

with $u(0,\cdot;u_0,c,d) = u_0(\cdot)$ $((c,d) \in H(a,b))$. By the comparison principle for parabolic equations [12], [34], Π_t is partially monotone with respect to the ordering $\leq$ in (2.2).

Definition 3.1. *A set $E \subset X \times H(a,b)$ is said to be trivial if $E \subset \{0\}\times H(a,b)$ and strictly positive if there is $u^+ \in \mathrm{Int}X_+$ such that $u \geq u^+$ for any $(u,c,d) \in E$.*

Notice that, for any $(c,d) \in H(a,b)$, $u \equiv 0$ is a solution of $(3.3)_{c,d}$ and that for sufficiently large $M > 1$, $u(t,x) = M$ is a supersolution of $(3.3)_{c,d}$. Therefore the comparison principle and standard a priori estimates for parabolic equations [12], [17], [18] yield the following result.

Lemma 3.2. *There is $M_0 > 0$ such that, for any $u_0 \in X_+ \setminus \{0\}$ and any $(c,d) \in H(a,b)$, $u(t,\cdot;u_0,c,d)$ exists for all $t \geq 0$ and*

$$u(t,\cdot;u_0,c,d) \gg 0 \quad \text{for} \quad t > 0,$$

$$0 < u(t,x;u_0,c,d) \leq M_0 \quad \text{for} \quad x \in \Omega, \quad t \gg 1.$$

Lemma 3.2 and the theory of parabolic equations [12], [17] imply that $\Pi_t : X_+ \times H(a,b) \to X_+ \times H(a,b)$ is a skew-product semiflow and $\{u(t,\cdot;u_0,c,d)|t \geq t_0\}$ is relatively compact in X for any $u_0 \in X_+$ and $t_0 > 0$. Therefore, the ω-limit set $\omega(u_0,c,d)$ of $\Pi_t(u_0,c,d)$ is well defined for all $(u_0,c,d) \in X_+ \times H(c,d)$. Moreover, Lemma 2.2(1) and the unique backward extensibility for parabolic equations show that Π_t has a flow extension on $\omega(u_0,c,d)$. The main result of this section is the following.

THEOREM 3.3. *Consider* (3.2). *One and only one of the following three alternatives occurs.*

(1) *For any $u_0 \in X_+ \setminus \{0\}$, $\omega(u_0,a,b)$ is strictly positive, a 1-cover of $H(a,b)$ (that is, $P^{-1}(c,d) \cap \omega(u_0,a,b)$ is a singleton for any $(c,d) \in H(a,b)$), and independent of u_0.*

(2) *For any $u_0 \in X_+ \setminus \{0\}$, $\omega(u_0,a,b) = \{0\} \times H(a,b)$.*

(3) *For any $u_0 \in X_+ \setminus \{0\}$, $\omega(u_0,a,b) \cap (\{0\} \times H(a,b)) \neq \emptyset$ and $\omega(u_0,a,b) \cap (\mathrm{Int}X_+ \times H(a,b)) \neq \emptyset$.*

The following corollary directly follows from Theorem 3.3 and Lemma 2.1.

COROLLARY 3.4. *Consider* (3.1). *One and only one of the following three alternatives occurs.*

(1) *There is a strictly positive almost periodic solution $u^*(t,x)$ with $\mathcal{M}(u^*) \subset \mathcal{M}(a,b)$ such that, for any $u_0 \in X_+ \setminus \{0\}$, $u(t,\cdot;u_0,a,b)$ converges to $u^*(t,x)$ in X; i.e.,*

$$\|u(t,\cdot;u_0,a,b) - u^*(t,\cdot)\| \to 0$$

as $t \to \infty$.

(2) *For any $u_0 \in X_+ \setminus \{0\}$, $u(t,\cdot;u_0,a,b)$ converges to $u \equiv 0$.*

(3) *For any $u_0 \in X_+ \setminus \{0\}$, $u(t,\cdot;u_0,a,b)$ is neither bounded away from $u \equiv 0$ nor converges to $u \equiv 0$.*

Before presenting the proof of Theorem 3.3, we establish the following four lemmas.

LEMMA 3.5. *Given $u_0 \in X_+$, $\alpha, \beta \in \mathbb{R}^+$ with $0 < \alpha \leq 1 \leq \beta$, and $(c,d) \in H(a,b)$, then*

$$\alpha u(t,\cdot;u_0,c,d) \leq u(t,\cdot;\alpha u_0,c,d) \leq u(t,\cdot;u_0,c,d)$$

and

$$u(t,\cdot;u_0,c,d) \leq u(t,\cdot;\beta u_0,c,d) \leq \beta u(t,\cdot;u_0,c,d)$$

for $t \geq 0$.

Proof. First of all, the comparison principle for parabolic equations yields

$$u(t,\cdot;\alpha u_0,c,d) \leq u(t,\cdot;u_0,c,d) \leq u(t,\cdot;\beta u_0,c,d) \tag{3.4}$$

for $t \geq 0$.

Next, letting $u_1(t,x) = \alpha u(t,x;u_0,c,d)$, $u_2(t,x) = u(t,x;\alpha u_0,c,d)$, $u_3(t,x) = u(t,x;\beta u_0,c,d)$, and $u_4(t,x) = \beta u(t,x;u_0,c,d)$, one has that $u = u_i(t,x)$ is the solution of

$$\begin{cases} u_t = k\Delta u + h(t,x)u, & x \in \Omega, \\ Bu = 0, & x \in \partial\Omega, \end{cases} \tag{3.5}$$

with $h(t,x) = h_i(t,x)$ and $u(0,x) = u_i(x)$ $(i = 1,2,3,4)$, where

$$h_1(t,x) = h_4(t,x) = c(t,x) - d(t,x)u(t,x;u_0,c,d),$$
$$h_2(t,x) = c(t,x) - d(t,x)u(t,x;\alpha u_0,c,d),$$
$$h_3(t,x) = c(t,x) - d(t,x)u(t,x;\beta u_0,c,d),$$

and

$$u_1(x) = u_2(x) = \alpha u_0(x), \qquad u_3(x) = u_4(x) = \beta u_0(x).$$

By (3.4),

$$h_1(t,x) \le h_2(t,x) \quad \text{and} \quad h_3(t,x) \le h_4(t,x) \quad \text{for} \quad x \in \Omega, \quad t \ge 0.$$

Again, it follows from the comparison principle that

$$u_1(t,\cdot) \le u_2(t,\cdot) \quad \text{and} \quad u_3(t,\cdot) \le u_4(t,\cdot) \tag{3.6}$$

for $t \ge 0$, and (3.4) and (3.6) imply the statements of the lemma. □

LEMMA 3.6. *Suppose that $\omega(u_0^*,a,b)$ is strictly positive for some $u_0^* \in \mathrm{Int}X_+$; then $\omega(u_0,a,b)$ is strictly positive for all $u_0 \in X_+ \setminus \{0\}$.*

Proof. First of all, $u(t,\cdot;u_0,a,b) \in \mathrm{Int}X_+$ for $u_0 \in X_+ \setminus \{0\}$ and $t > 0$. Hence we may assume without loss of generality that $u_0 \in IntX_+$.

Next, for any $u_0 \in \mathrm{Int}X_+$, there is $\alpha > 0$ such that $u_0 \ge \alpha u_0^*$. Then the comparison principle for parabolic equations and Lemma 3.5 imply

$$\alpha u(t,x;u_0^*,a,b) \le u(t,\cdot;u_0,a,b) \quad \text{for} \quad t > 0.$$

Thus, $\omega(u_0,a,b)$ is strictly positive for any $u_0 \in \mathrm{Int}X_+$. □

LEMMA 3.7. *Let $E \subset X_+ \times H(a,b)$ be compact, strictly positive, and invariant under Π_t. Then, for each $\epsilon > 0$, there is $\delta > 0$ such that, for any $(u_1,c,d) \in E$ and $u_2 \in X_+$ with $\|u_2 - u_1\| < \delta$,*

$$|u(t,x;u_2,c,d) - u(t,x;u_1,c,d)| < \epsilon \quad \text{for} \quad x \in \Omega, \quad t \ge 0.$$

Proof. Since E is compact, strictly positive, and invariant under Π_t, there exists a $u_+ \in \mathrm{Int}X_+$ such that $u \ge u_+$ for all $(u,c,d) \in E$. Also, one finds an $M > 0$ such that $|u(x)| \le M$ for every $(u,c,d) \in E$ and $x \in \Omega$. Now let $\epsilon > 0$ and select $0 < \alpha < 1$ with $\alpha M < \epsilon$. We claim that there is $\delta > 0$ such that, for each $(u_1,c,d) \in E$ and $u_2 \in X_+$ with $\|u_2 - u_1\| < \delta$,

$$(1-\alpha)u_1(\cdot) \le u_2(\cdot) \le (1+\alpha)u_1(\cdot). \tag{3.7}$$

Otherwise, let $\delta_n = \frac{1}{n}$. Then there exist $(u_n^1,c_n,d_n) \in E$ and $u_n^2 \in X_+$ with $\|u_n^2 - u_n^1\| < \delta_n$, but (3.7) does not hold for $u_1 = u_n^1$ and $u_2 = u_n^2$. Without loss of generality, assume that $(u_n^1,c_n,d_n) \to (u_1^*,c^*,d^*)$ as $n \to \infty$. We have

$$\frac{1+\alpha/2}{1+\alpha}u_1^* \le u_n^1 \le \frac{1-\alpha/2}{1-\alpha}u_1^*$$

and

$$(1-\alpha/2)u_1^* \le u_n^2 \le (1+\alpha/2)u_1^*$$

for n large enough. This implies that

$$(1-\alpha)u_n^1 \le (1-\alpha/2)u_1^* \le u_n^2 \le (1+\alpha/2)u_1^* \le (1+\alpha)u_n^1$$

for n large enough, a contradiction. Hence (3.7) holds for some $\delta > 0$. Then the comparison principle for parabolic equations and Lemma 3.5 show that

$$(1-\alpha)u(t,\cdot;u_1,c,d) \le u(t,\cdot;u_2,c,d) \le (1+\alpha)u(t,\cdot;u_1,c,d)$$

for all $t \ge 0$ and $(u_1,c,d) \in E$, $u_2 \in X_+$ with $\|u_2 - u_1\| < \delta$, hence

$$|u(t,x;u_2,c,d) - u(t,x;u_1,c,d)| \le \alpha|u(t,x;u_1,c,d)| \le \alpha M < \epsilon$$

for $x \in \Omega$ and $t \ge 0$. □

LEMMA 3.8. *If $\omega(u_0,a,b)$ is strictly positive for $u_0 \in X_+ \setminus \{0\}$, then it is independent of u_0.*

Proof. It is sufficient to prove that, for any $u_1, u_2 \in \mathrm{Int}X_+$ with $u_1 \ll u_2$, $\omega(u_1,a,b) = \omega(u_2,a,b)$.

To this end, fix $u_1^0, u_2^0 \in \mathrm{Int}X_+$ with $u_1^0 \ll u_2^0$. By Lemma 3.6, both $\omega(u_1^0,a,b)$ and $\omega(u_2^0,a,b)$ are strictly positive. Suppose that $\omega(u_1^0,a,b) \ne \omega(u_2^0,a,b)$. Then we may assume that there is $(u_1^*,c,d) \in \omega(u_1^0,a,b) \setminus \omega(u_2^0,a,b)$. Let $t_n \to \infty$ and $(u_2^*,c,d) \in \omega(u_2^0,c,d)$ be such that $\Pi_{t_n}(u_1^0,a,b) \to (u_1^*,c,d)$ and $\Pi_{t_n}(u_2^0,a,b) \to (u_2^*,c,d)$ as $n \to \infty$. By $u_1^0 \ll u_2^0$,

$$u(t,\cdot;u_1^*,c,d) = \lim_{n\to\infty} u(t+t_n,\cdot;u_1^0,a,b) \le \lim_{n\to\infty} u(t,\cdot;u_2^0,a,b) = u(t,\cdot;u_2^*,c,d)$$

for $t \in \mathbb{R}$. Clearly, $u_1^* \ne u_2^*$, and the comparison principle yields

$$u(t,\cdot;u_1^*,c,d) \ll u(t,\cdot;u_2^*,c,d) \quad \text{for} \quad t \in \mathbb{R}.$$

Moreover, we claim that there is $u_+(\cdot) \in \mathrm{Int}X_+$ such that

$$u(t,\cdot;u_2^*,c,d) - u(t,\cdot;u_1^*,c,d) \ge u_+(\cdot) \quad \text{for} \quad t \le 0. \tag{3.8}$$

Otherwise, there are $u_n^+(\cdot) \in \mathrm{Int}X_+$ with $\|u_n^+\| \to 0$ as $n \to \infty$ and $t_n \le 0$ such that (3.8) does not hold for $u_+ = u_n^+$ and $t = t_n$. Without loss of generality, assume that $u(t_n,\cdot;u_i^*,c,d) \to u_i^{**}(\cdot)$ and $(c,d)\cdot t_n \to (c^*,d^*)$. By Lemma 3.7, $u_1^{**} \ne u_2^{**}$, thus $u_1^{**} \ll u_2^{**}$ thanks to the comparison principle for parabolic equations. Hence there is $u_+^*(\cdot) \in \mathrm{Int}X_+$ such that

$$u_2^{**}(\cdot) - u_1^{**}(\cdot) \gg u_+^*(\cdot)$$

and then

$$u(t_n,\cdot;u_2^*,c,d) - u(t_n,\cdot;u_1^*,c,d) \ge \frac{1}{2}u_+^*(\cdot) \ge u_n^+(\cdot)$$

for n large enough, a contradiction. Therefore, (3.8) holds for some $u_+(\cdot) \in \mathrm{Int}X_+$.

Next, select $s_n \to -\infty$ such that

$$u(s_n,\cdot;u_i^*,c,d) \to \tilde{u}_i^*(\cdot) \quad \text{and} \quad (c,d)\cdot s_n \to (a,b),$$

and let $u_i(t,x) = u(t,x;\tilde{u}_i^*,a,b)$ $(i=1,2)$. By (3.8),

$$u_2(t,\cdot) - u_1(t,\cdot) \ge u_+(\cdot) \quad \text{for} \quad t \in \mathbb{R}. \tag{3.9}$$

By the positivity and compactness of $\omega(u_i^0, a, b)$ $(i = 1, 2)$, there is $\alpha > 0$ such that

$$u_2(t, \cdot) \le \alpha u_1(t, \cdot) \quad \text{for} \quad t \in \mathbb{R}. \tag{3.10}$$

Now let $\Phi^1(t, x)$ be the evolution operator of (3.5) with $h(t, x) = a(t, x) - b(t, x)u_1(t, x)$; that is, $\Phi^1(t, s)u_0 = u(t; s, u_0)$, where $u(t; s, u_0)$ is the solution of (3.5) with $h(t, x) = a(t, x) - b(t, x)u_1(t, x)$ and $u(s; s, u_0) = u_0$. We claim that there is $\beta > 0$ such that

$$\Phi^1(s + 1, s)\big(b(s, \cdot)(u_2(s, \cdot) - u_1(s, \cdot))u_2(s, \cdot)\big) \ge \beta u_1(s + 1, \cdot) \tag{3.11}$$

for any $s \in \mathbb{R}$. In fact, otherwise one finds for $\beta_n = \frac{1}{n}$ an $s_n \in \mathbb{R}$ such that

$$\Phi^1(s_n + 1, s_n)\big(b(s_n, \cdot)(u_2(s_n, \cdot) - u_1(s_n, \cdot))u_2(s_n, \cdot)\big) \not\ge \beta_n u_1(s_n + 1, \cdot). \tag{3.12}$$

Without loss of generality, we may assume that

$$u_i(s_n, \cdot) \to \bar{u}_i(\cdot), \qquad (a \cdot s_n, b \cdot s_n) \to (\bar{a}, \bar{b})$$

as $n \to \infty$. Let $\bar{\Phi}^1(t, s)$ be the evolution operator of (3.5) with $h(t, x) = \bar{a}(t, x) - \bar{b}(t, x)\bar{u}_1(t, x)$, where $\bar{u}_1(t, x) = \lim_{n\to\infty} u_1(t + s_n, x)$. Then

$$\begin{aligned}\Phi^1(s_n + 1, s_n)\big(b(s_n, \cdot)(u_2(s_n, \cdot) - u_1(s_n, \cdot))u_2(s_n, \cdot)\big)\\ \to \bar{\Phi}^1(1, 0)\big(\bar{b}(0, \cdot)(\bar{u}_2(\cdot) - \bar{u}_1(\cdot))\bar{u}_2(\cdot)\big)\end{aligned}$$

as $n \to \infty$. By (3.9) and the positivity of $\omega(u_2^0, a, b)$,

$$\bar{\Phi}^1(1, 0)\big(\bar{b}(0, \cdot)(\bar{u}_2(\cdot) - \bar{u}_1(\cdot))\bar{u}_2(\cdot)\big) \gg 0.$$

Hence there is $\bar{\beta} > 0$ such that

$$\bar{\Phi}^1(1, 0)\big(\bar{b}(0, \cdot)(\bar{u}_2(\cdot) - \bar{u}_1(\cdot))\bar{u}_2(\cdot)\big) \gg 2\bar{\beta}\bar{u}_1(1, \cdot).$$

This implies that

$$\Phi^1(s_n + 1, s_n)\big(b(s_n, \cdot)(u_2(s_n, \cdot) - u_1(s_n, \cdot))u_2(s_n, \cdot)\big) \gg \bar{\beta}\bar{u}_1(1, \cdot)$$

for n large enough; consequently, there is $\beta > 0$ such that

$$\Phi^1(s_n + 1, s_n)\big(b(s_n, \cdot)(u_2(s_n, \cdot) - u_1(s_n, \cdot))u_2(s_n, \cdot)\big) \ge \beta u_1(s_n + 1, \cdot)$$

for n large enough, a contradiction to (3.12). Hence there is $\beta > 0$ such that (3.11) holds.

Note that $u_2(t, x)$ is a solution of (3.5) with $h(t, x) = a(t, x) - b(t, x)u_2(t, x) = a(t, x) - b(t, x)u_1(t, x) - b(t, x)\big(u_2(t, x) - u_1(t, x)\big)$. By the variation of constant formula [17] and (3.10), (3.11), for $t \ge 1$,

$$\begin{aligned}u_2(t, x) &= \Phi^1(t, s)u_2(s, \cdot) - \int_0^t \Phi^1(t, s)\big(b(s, \cdot)(u_2(s, \cdot) - u_1(s, \cdot))u_2(s, \cdot)\big)ds\\
&\le \Phi^1(t, s)\alpha u_1(s, \cdot) - \int_0^{t-1} \Phi^1(t, s + 1)\Phi^1(s + 1, s)\big(b(s, \cdot)(u_2(s, \cdot) - u_1(s, \cdot))\\
&\qquad \times u_2(s, \cdot)\big)ds\\
&\le \alpha u_1(t, \cdot) - \int_0^{t-1} \Phi^1(t, s + 1)\beta u_1(s + 1, \cdot)ds\\
&= \alpha u_1(t, \cdot) - \beta \int_0^{t-1} u_1(t, \cdot)ds\\
&= (\alpha - \beta(t - 1))u_1(t, \cdot),\end{aligned}$$

which contradicts the boundedness and positivity of u_1, u_2. Hence $\omega(u_1^0, a, b) = \omega(u_2^0, a, b)$. □

Proof of Theorem 3.3. First of all, fix a $u_0^* \in \mathrm{Int}X_+$. Then $\omega(u_0^*, a, b)$ is either strictly positive, or trivial, or neither strictly positive nor trivial.

Claim 1. $\omega(u_0^*, a, b)$ is strictly positive iff alternative (1) of Theorem 3.3 occurs.

First, if alternative (1) of Theorem 3.3 occurs, clearly, then $\omega(u_0^*, a, b)$ is strictly positive.

Next, suppose that $\omega(u_0^*, a, b)$ is strictly positive. Then, by Lemmas 3.6 and 3.8, $\omega(u_0, a, b)$ is strictly positive and independent of u_0 for any $u_0 \in \mathrm{Int}X_+$. It remains to prove that $\omega(u_0^*, a, b)$ is a 1-cover of $H(a, b)$. Note that there is $u_+ \in \mathrm{Int}X_+$ such that $u(\cdot) \le u_+(\cdot)$ for any $(u, a, b) \in \omega(u_0^*, a, b)$. Hence $u \le u_+$ for any $(u, a, b) \in \omega(u_+, a, b)$. Then by Lemma 2.2(3), $\omega(u_0^*, a, b) = \omega(u_+, a, b)$ is an almost 1-cover of $H(a, b)$. Suppose that $(c_0, d_0) \in H(a, b)$ is such that $\omega(u_0^*, a, b) \cap (X \times \{(c_0, d_0)\}) = \{(u_0, c_0, d_0)\}$ is a singleton. Then one obtains for every (u_1, c, d), $(u_2, c, d) \in \omega(u_0^*, a, b)$ and $s_n \to -\infty$ with $(c, d) \cdot s_n \to (c_0, d_0)$ that

$$\|u(s_n, \cdot; u_1, c, d) - u(s_n, \cdot; u_2, c, d)\| \to 0$$

as $n \to \infty$. By Lemma 3.7, we must have $u_1 = u_2$ and then $\omega(u_0^*, a, b)$ is a 1-cover of $H(a, b)$.

Claim 2. $\omega(u_0^*, a, b)$ is trivial iff alternative (2) of Theorem 3.3 occurs.

First, observe that, if alternative (2) of Theorem 3.3 occurs, then $\omega(u_0^*, a, b)$ is trivial.

Next, suppose that $\omega(u_0^*, a, b)$ is trivial. Note that $u(t, \cdot; u_0, a, b) \gg 0$ for all $t > 0$ and $u_0 \in X_+ \setminus \{0\}$. Hence we need only to prove that $\omega(u_0, a, b)$ is trivial for each $u_0 \in \mathrm{Int}X_+$. Given any $u_0 \in \mathrm{Int}X_+$, let $\alpha > 0$ be such that $\alpha u_0(\cdot) \le u_0^*(\cdot)$. By Lemma 3.5,

$$\alpha u(t, x; u_0, a, b) \le u(t, x; u_0^*, a, b)$$

for $x \in \Omega$ and $t \ge 0$. This implies that $\omega(u_0, a, b)$ is trivial, and hence the claim follows.

Claim 3. $\omega(u_0^*, a, b)$ is neither trivial nor strictly positive iff alternative (3) of Theorem 3.3 occurs.

Clearly, if alternative (3) of Theorem 3.3 holds, then $\omega(u_0^*, a, b)$ is neither trivial nor strictly positive.

Conversely, suppose that $\omega(u_0^*, a, b)$ is neither trivial nor strictly positive. Then we also have $\omega(u_0, a, b)$ is neither trivial nor strictly positive for any $u_0 \in X_+ \setminus \{0\}$. Since $\omega(u_0, a, b) \ne \{0\} \times H(a, b)$, there is $(u^*, c, d) \in \omega(u_0, a, b)$ with $u^* \in X_+ \setminus \{0\}$. Then we must have $u(t, \cdot; u^*, c, d) \in \mathrm{Int}X_+$ for $t > 0$. Hence $\omega(u_0, a, b) \cap (\mathrm{Int}X_+ \times H(a, b)) \ne \emptyset$. Since $\omega(u_0, a, b)$ is not strictly positive, for any $u_n^+ \in \mathrm{Int}X_+$ with $\|u_n^+\| \to 0$, there is t_n such that $u(t_n, \cdot; u_0, a, b) \not\ge u_n^+(\cdot)$. Assume that $u(t_n, \cdot; u_0, a, b) \to u^*(\cdot)$ as $n \to \infty$. Then we must have $u^* = 0$. Hence $\omega(u_0, a, b) \cap (\{0\} \times H(a, b)) \ne \emptyset$.

Theorem 3.3 now follows from Claims 1, 2, and 3. □

Remark 3.1. (1) When a, b are actually periodic, alternative (3) of Theorem 3.3 does not occur (cf. [19], [21]).

(2) In the case where a is almost periodic and homogeneous Neumann boundary conditions are prescribed, the theorem has been proved in [36] by means of a very different approach which is limited to Neumann conditions. Moreover, it has been shown in [36] that each of the alternatives (1), (2), (3) really occurs.

COROLLARY 3.9. *Let* $\bar{a}(x) = \lim_{t\to\infty} \frac{1}{t}\int_0^t a(s,x)ds$. *If* $\lambda(k,a) > 0$ *or* $\lambda(k,\bar{a}) > 0$, *then alternative* (1) *of Theorem* 3.3 *occurs.*

Proof. By Lemma 2.3(4), there is $\sigma_1 < \lambda(k,a)$ such that $\Sigma(k,H(a)) = \Sigma_1 \cup \{\lambda(k,a)\}$ and $\lambda \le \sigma_1$ for any $\lambda \in \Sigma_1$. Hence $u \equiv 0$ is a linearly unstable solution of (3.1), and the invariant manifold theory ([5], [7], etc.) excludes alternatives (2) and (3) of Theorem 3.3, thus alternative (1) holds. □

4. Basic properties of almost periodic two species competition models. In this section, we present some basic properties of the two species competition model:

$$(4.1)\qquad \begin{cases} u_t = k_1\Delta u + u(a_1(t,x) - b_1(t,x)u - c_1(t,x)v), & x \in \Omega,\\ v_t = k_2\Delta v + v(a_2(t,x) - b_2(t,x)u - c_2(t,x)v), & x \in \Omega,\\ Bu = Bv = 0, & x \in \partial\Omega, \end{cases}$$

where k_i, a_i, b_i, c_i $(i = 1,2)$, Ω and Bu, Bv are as in (1.1), and b_1, $c_2 \ge \delta$ for some $\delta > 0$, c_1, $b_2 \ge 0$. Let $f_i(t,x,u,v) = a_i(t,x) - b_i(t,x)u - c_i(t,x)v$ $(i = 1,2)$. Let X be as in section 2.3 and $\Pi_t : X \times X \times H(f_1,f_2) \to X \times X \times H(f_1,f_2)$ be the (local) skew-product semiflow generated by (4.1),

$$(4.2)\qquad \Pi_t(u_0,v_0,g_1,g_2) = (u(t,\cdot;u_0,v_0,g_1,g_2), v(t,\cdot;u_0,v_0,g_1,g_2), g_1\cdot t, g_2\cdot t),$$

where $(u(t,\cdot;u_0,v_0,g_1,g_2), v(t,\cdot;u_0,v_0,g_1,g_2))$ is the solution of

$$(4.3)_{g_1,g_2}\qquad \begin{cases} u_t = k_1\Delta u + ug_1(t,x,u,v), & x \in \Omega,\\ v_t = k_2\Delta v + vg_2(t,x,u,v), & x \in \Omega,\\ Bu = Bv = 0, & x \in \partial\Omega \end{cases}$$

with $(u_0(\cdot), v_0(\cdot)) := (u(0,\cdot;u_0,v_0,g_1,g_2), v(0,\cdot;u_0,v_0,g_1,g_2))$.

Given (u_1,v_1), $(u_2,v_2) \in X_+ \times X_+$, we define

$$(4.4)_1\qquad (u_1,v_1) \le_1 (u_2,v_2) \quad\text{if}\quad u_1 \le u_2, \quad v_1 \le v_2,$$

$$(4.4)_2\qquad (u_1,v_1) <_1 (u_2,v_2) \quad\text{if}\quad (u_1,v_1) \le_1 (u_2,v_2), \quad (u_1,v_1) \ne (u_2,v_2),$$

$$(4.4)_3\qquad (u_1,v_1) \ll_1 (u_2,v_2) \quad\text{if}\quad (u_2-u_1, v_2-v_1) \in \mathrm{Int}(X_+) \times \mathrm{Int}(X_+),$$

$$(4.5)_1\qquad (u_1,v_1) \le_2 (u_2,v_2) \quad\text{if}\quad u_1 \le u_2, \quad v_1 \ge v_2,$$

$$(4.5)_2\qquad (u_1,v_1) <_2 (u_2,v_2) \quad\text{if}\quad (u_1,v_1) \le_2 (u_2,v_2), \quad (u_1,v_1) \ne (u_2,v_2),$$

$$(4.5)_3\qquad (u_1,v_1) \ll_2 (u_2,v_2) \quad\text{if}\quad (u_2-u_1, v_1-v_2) \in \mathrm{Int}(X_+) \times \mathrm{Int}(X_+).$$

For (u_1,v_1,g_1,g_2), $(u_2,v_2,g_1,g_2) \in X_+ \times X_+ \times H(f_1,f_2)$, define

$$(4.6)_1\quad (u_1,u_2,g_1,g_2) \le_1 (<_1,\ll_1)(u_2,v_2,g_1,g_2) \quad\text{if}\quad (u_1,v_1) \le_1 (<_1,\ll_1)(u_2,v_2)$$

and

$$(4.6)_2\quad (u_1,u_2,g_1,g_2) \le_2 (<_2,\ll_2)(u_2,v_2,g_1,g_2) \quad\text{if}\quad (u_1,v_1) \le_2 (<_2,\ll_2)(u_2,v_2).$$

LEMMA 4.1. (1)

$$\Pi_t(X_+ \times \{0\} \times H(f_1,f_2)) \subset X_+ \times \{0\} \times H(f_1,f_2) \quad\text{for}\quad t > 0,$$

$$\Pi_t(\{0\} \times X_+ \times H(f_1,f_2)) \subset \{0\} \times X_+ \times H(f_1,f_2) \quad\text{for}\quad t > 0.$$

(2)

$$\Pi_t(X_+ \times X_+ \times H(f_1, f_2)) \subset X_+ \times X_+ \times H(f_1, f_2) \quad \text{for} \quad t > 0.$$

Proof. These results follow from the standard parabolic theory. □

By Lemma 4.1, $\Pi_t|_{X_+\times X_+\times H(f_1,f_2)}$ ($\Pi|_{X_+\times\{0\}\times H(f_1,f_2)}$, $\Pi|_{\{0\}\times X_+\times H(f_1,f_2)}$) is a skew-product semiflow.

LEMMA 4.2. *If* (u_1, v_1), $(u_2, v_2) \in X_+ \times X_+$ *and* $(u_1, v_1) \leq_2 (u_2, v_2)$, *then*

$$\Pi_t(u_1, v_1, g_1, g_2) \leq_2 \Pi_t(u_2, v_2, g_1, g_2)$$

for all $t > 0$ *and* $(g_1, g_2) \in H(f_1, f_2)$. *Moreover, if* $(u_1, v_1) <_2 (u_2, v_2)$ *and* $(u_1, v_1) \notin X_+ \times \{0\}$, $(u_2, v_2) \notin \{0\} \times X_+$, *then*

$$\Pi_t(u_1, v_1, g_1, g_2) \ll_2 \Pi_t(u_2, v_2, g_1, g_2)$$

for all $t > 0$ *and* $(g_1, g_2) \in H(f_1, f_2)$.

Proof. The proof follows from Lemma 4.1 and the comparison principle for parabolic equations. □

By Lemmas 4.1 and 4.2, $\Pi_t|_{X_+\times X_+\times H(f_1,f_2)}$ is partially monotone with respect to the ordering $\leq_2$ in $(4.5)_1$, and $\Pi_t(u_0, v_0, g_1, g_2)$ is strictly positive for all $(u_0, v_0, g_1, g_2) \in (X_+ \setminus \{0\}) \times (X_+ \setminus \{0\}) \times H(f_1, f_2)$ and $t > 0$; that is,

$$(0, 0, g_1 \cdot t, g_2 \cdot t) \ll_1 \Pi_t(u_0, v_0, g_1, g_2)$$

for $t > 0$.

LEMMA 4.3. *Assume that* $\lambda(k_1, a_1) > 0$ *and* $\lambda(k_2, a_2) > 0$.

(1) *There is* $E_1 \subset X_+ \times \{0\} \times H(f_1, f_2)$, *which is invariant under* Π_t *and has the form*

$$E_1 = \{(u_{g_1}, 0, g_1, g_2) | (g_1, g_2) \in H(f_1, f_2)\}.$$

E_1 *is attracting in the sense that, for each* $u_0 \in X_+ \setminus \{0\}$ *and* $(g_1, g_2) \in H(f_1, f_2)$,

$$\|u(t, \cdot; u_0, 0, g_1, g_0) - u(t, \cdot; u_{g_1}, 0, g_1, g_2)\| \to 0$$

as $t \to \infty$.

(2) *There is* $E_2 \subset \{0\} \times X_+ \times H(f_1, f_2)$, *which is invariant under* Π_t *and has the form*

$$E_2 = \{(0, v_{g_2}, g_1, g_2) | (g_1, g_2) \in H(f_1, f_2)\}.$$

E_2 *is attracting in the sense that, for each* $v_0 \in X_+ \setminus \{0\}$ *and* $(g_1, g_2) \in H(f_1, f_2)$,

$$\|u(t, \cdot; 0, v_0, g_1, g_2) - u(t, \cdot; 0, v_{g_2}, g_1, g_2)\| \to 0$$

as $t \to \infty$.

Proof. The proof follows from Corollary 3.9 and Lemma 4.1(1). □

Unless otherwise specified, we assume throughout the rest of the paper that $\lambda(k_1, a_1) > 0$ and $\lambda(k_2, a_2) > 0$. Let $E \subset X_+ \times X_+ \times H(f_1, f_2)$ be such that

$$E \cap (X \times X \times \{(g_1, g_2)\}) = ([0, u_{g_1}] \times [0, v_{g_2}] \times \{(g_1, g_2)\}) \setminus \{(0, 0, g_1, g_2)\}, \tag{4.7}$$

where $[0, u_{g_1}] = \{u \in X \mid 0 \leq u \leq u_{g_1}\}$ and $[0, v_{g_2}] = \{v \in X \mid 0 \leq v \leq v_{g_2}\}$.

LEMMA 4.4. (1) $\Pi_t E \subset E$ *for each* $t > 0$*;*

(2) $\omega(u_0, v_0, g_1, g_2) \subset E$ *for all* $(u_0, v_0) \in (X_+ \times X_+) \setminus \{(0,0)\}$ *and* $(g_1, g_2) \in H(f_1, f_2)$.

Proof. The proof follows from Lemmas 4.2 and 4.3. □

Let $a_{iL(M)}$, $b_{iL(M)}$, and $c_{iL(M)}$ have the same meaning as in $(1.4)_1$–$(1.4)_3$, and let $(u^{\pm}(t,\cdot;u_0,v_0), v^{\pm}(t,\cdot;u_0,v_0))$ denote the solutions of

$$(4.8)_+ \qquad \begin{cases} u_t = k_1\Delta u + u(a_{1M} - b_{1L}u - c_{1L}v), & x \in \Omega, \\ v_t = k_2\Delta v + v(a_{2L} - b_{2M}u - c_{2M}v), & x \in \Omega, \\ Bu = Bv = 0, & x \in \partial\Omega \end{cases}$$

(this is the case +) and

$$(4.8)_- \qquad \begin{cases} u_t = k_1\Delta u + u(a_{1L} - b_{1M}u - c_{1M}v), & x \in \Omega, \\ v_t = k_2\Delta v + v(a_{2M} - b_{2L}u - c_{2L}v), & x \in \Omega, \\ Bu = Bv = 0, & x \in \partial\Omega \end{cases}$$

satisfying $(u^{\pm}(0,\cdot;u_0,v_0), v^{\pm}(0,\cdot;u_0,v_0)) = (u_0(\cdot), v_0(\cdot))$. The comparison principle for parabolic equations yields the following result.

LEMMA 4.5. *If* $(u_0, v_0) \in X_+ \times X_+$ *and* $(g_1, g_2) \in H(f_1, f_2)$*, then*

$$(u^-(t,\cdot;u_0,v_0), v^-(t,\cdot;u_0,v_0)) \leq_2 (u(t,\cdot;u_0,v_0,g_1,g_2), v(t,\cdot;u_0,v_0,g_1,g_2))$$

and

$$(u(t,\cdot;u_0,v_0,g_1,g_2), v(t,\cdot;u_0,v_0,g_1,g_2)) \leq_2 (u^+(t,\cdot;u_0,v_0), v^+(t,\cdot;u_0,v_0))$$

for $t \geq 0$.

The following lemmas concern the dynamics of (4.1) when a_i, b_i, and c_i $(i = 1, 2)$ are constants.

LEMMA 4.6. *Assume that* a_i, b_i, c_i $(i = 1, 2)$ *are positive constants and that* $\lambda(k_1, a_1) > 0$, $\lambda(k_2, a_2) > 0$. *Then exactly one of the following alternatives holds.*

(1) (4.1) *has a strictly positive equilibrium solution.*

(2) *Every positive solution of* (4.1) *converges to the solution* $(u_{f_1}, 0)$.

(3) *Every positive solution of* (4.1) *converges to the solution* $(0, v_{f_2})$.

Proof. See [25]. □

LEMMA 4.7. *Assume that* a_i, b_i, c_i $(i = 1, 2)$ *are positive constants and* $Bu = \frac{\partial u}{\partial n}$.

(1) *If* $a_1 > \frac{c_1 a_2}{c_2}$ *and* $a_2 > \frac{a_1 b_2}{b_1}$, *then every positive solution of* (4.1) *converges to a unique strictly positive equilibrium.*

(2) *If* $a_1 > \frac{c_1 a_2}{c_2}$ *and* $a_2 \leq \frac{a_1 b_2}{b_1}$, *then every positive solution of* (4.1) *converges to the equilibrium* $(u_{f_1}, 0) \equiv (\frac{a_1}{b_1}, 0)$.

(3) *If* $a_1 \leq \frac{c_1 a_2}{c_2}$ *and* $a_2 > \frac{a_1 b_2}{b_1}$, *then every positive solution of* (4.1) *converges to the equilibrium* $(0, v_{f_2}) \equiv (0, \frac{a_2}{c_2})$.

Proof. (1) The proof follows from [31], [48].

(2) and (3) follow from [3], [31]. □

LEMMA 4.8. *Assume that* a_i, b_i, c_i $(i = 1, 2)$ *are positive constants,* $Bu = u$, *and* $\lambda(k_1, a_1) > 0$, $\lambda(k_2, a_2) > 0$.

(1) *If* $a_1 > \frac{c_1 a_2}{c_2}$, $a_2 > \frac{a_1 b_2}{b_1}$, $k_1 = k_2$, *and* $a_1 = a_2$, *then every positive solution of* (4.1) *converges to a unique strictly positive equilibrium.*

(2) *If* $a_1 > \frac{c_1 a_2}{c_2}$, $a_2 \leq \frac{a_1 b_2}{b_1}$, $k_1 \leq k_2$, *and* $a_1 \geq a_2$, *then every positive solution of* (4.1) *converges to* $(u_{f_1}, 0)$.

(3) *If* $a_1 \le \frac{c_1 a_2}{c_2}$, $a_2 > \frac{a_1 b_2}{b_1}$, $k_1 \ge k_2$, *and* $a_1 \le a_2$, *then every positive solution of* (4.1) *converges to* $(0, v_{f_2})$.

Proof. (1) The proof follows from [8].

(2) First, we show that $(0, v_{f_2})$ is linearly unstable. Since v_{f_2} is a solution of (3.5) with $k = k_2$ and $h = a_2 - c_2 v_{f_2}$, one has $\lambda(k_2, a_2 - c_2 v_{f_2}) = 0$. Linearizing (4.1) around $(0, v_{f_2})$, we obtain

$$\begin{cases} u_t = k_1 \Delta u + (a_1 - c_1 v_{f_2})u, & x \in \Omega, \\ v_t = k_2 \Delta v - b_2 v_{f_2} u + (a_2 - 2c_2 v_{f_2})v, & x \in \Omega, \\ u = v = 0, & x \in \partial\Omega. \end{cases} \tag{4.9}$$

Note that $0 < v_{f_2} \le \frac{a_2}{c_2}$; consequently, $a_1 \ge a_2$ and $a_1 > \frac{c_1 a_2}{c_2}$ yield

$$a_1 - c_1 v_{f_2} = a_1 \left(1 - \frac{c_1}{a_1} v_{f_2}\right) > a_1 \left(1 - \frac{c_2}{a_2} v_{f_2}\right) \ge a_2 \left(1 - \frac{c_2}{a_2} v_{f_2}\right) = a_2 - c_2 v_{f_2}.$$

Therefore $k_1 \le k_2$ and Lemma 2.3(5) imply $\lambda(k_1, a_1 - c_1 v_{f_2}) > 0$. Observe that $\lambda = \lambda(k_1, a_1 - c_1 v_{f_2})$ is an eigenvalue of the following eigenvalue problem which arises from (4.9):

$$\begin{cases} k_1 \Delta u + (a_1 - c_1 v_{f_2})u = \lambda u, & x \in \Omega, \\ k_2 \Delta v - b_2 v_{f_2} u + (a_2 - 2c_2 v_{f_2})v = \lambda v, & x \in \Omega, \\ u = v = 0, & x \in \partial\Omega. \end{cases} \tag{4.10}$$

Hence $(0, v_{f_2})$ is linearly unstable, and therefore alternative (3) of Lemma 4.6 cannot occur.

Next, we prove (4.1) has no strictly positive equilibrium. Suppose that there is a strictly positive equilibrium (u_*, v_*). Then $u = u_*$ is a solution of (3.5) with $k = k_1$ and $h = a_1 - b_1 u_* - c_1 v_*$, and $v = v_*$ is a solution of (3.5) with $k = k_2$ and $h = a_2 - b_2 u_* - c_2 v_*$. Hence $\lambda(k_1, a_1 - b_1 u_* - c_1 v_*) = 0$ and $\lambda(k_2, a_2 - b_2 u_* - c_2 v_*) = 0$. By $a_1 \ge a_2$, $a_1 > \frac{c_1 a_2}{c_2}$, and $a_2 \le \frac{a_1 b_2}{b_1}$, one gets

$$a_1 - b_1 u_* - c_1 v_* > a_1 - \frac{a_1 b_2}{a_2} u_* - \frac{a_1 c_2}{a_2} v_* = \frac{a_1}{a_2}(a_2 - b_2 u_* - c_2 v_*). \tag{4.11}$$

Let $\xi_2(x)$ ($\|\xi_2\|_2 = 1$) be a positive eigenfunction of (2.9) with $k = k_2$, $h = a_2 - b_2 u_* - c_2 v_*$, and $\lambda = \lambda(k_2, a_2 - b_2 u_* - c_2 v_*)$. Then

$$k_2 \int_\Omega |\bigtriangledown \xi_2|^2 dx = \int_\Omega (a_2 - b_2 u_* - c_2 v_*)\xi_2^2 dx > 0. \tag{4.12}$$

By (4.11) and (4.12),

$$\begin{aligned} -k_1 \int_\Omega |\bigtriangledown \xi_2|^2 dx + \int_\Omega (a_1 - b_1 u_* - c_1 v_*)\xi_2^2 dx \\ > -k_2 \int_\Omega |\bigtriangledown \xi_2|^2 dx + \frac{a_1}{a_2} \int_\Omega (a_2 - b_2 - c_2 v_*)\xi_2^2 dx \ge 0. \end{aligned}$$

It then follows from the arguments of Lemma 2.3(5) that $\lambda(k_1, a_1 - b_1 u_* - c_1 v_*) > 0$, a contradiction. Therefore, (4.1) has no strictly positive equilibrium.

Now, by Lemma 4.6, every positive solution of (4.1) converges to $(u_{f_1}, 0)$.

(3) The proof can be derived by arguments similar to those in (2). □

5. Uniform persistence, coexistence, and extinction in almost periodic two species competition models. Let $a_{iL(M)}$, $b_{iL(M)}$, $c_{iL(M)}$ have the same meaning as in $(1.4)_1$–$(1.4)_3$, and let u_{g_1}, v_{g_2} be understood as in Lemma 4.3. Then we obtain for the Neumann case the following result.

THEOREM 5.1. *Consider* (4.1). *Suppose that* $Bu = \frac{\partial u}{\partial n}$ *and* a_{iL}, b_{iL}, *and* c_{iL} $(i = 1, 2)$ *are positive.*

(1) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$ *and* $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, *then there exist* (u_-, v_-), $(u_+, v_+) \in \mathrm{Int}X_+ \times \mathrm{Int}X_+$ *with* $(u_-, v_-) \ll_2 (u_+, v_+)$ *such that, for each* $(u_0, v_0) \in (X_+ \setminus \{0\}) \times (X_+ \setminus \{0\})$ *and each* $(u, v, g_1, g_2) \in \omega(u_0, v_0, f_1, f_2)$,

$$(u_-, v_-) \leq_2 (u, v) \leq_2 (u_+, v_+)$$

(hence uniform persistence occurs). Moreover, there are (u_*^-, v_*^-), (u_*^+, v_*^+) *with*

$$(u_-, v_-) \leq_2 (u_*^-, v_*^-) \leq_2 (u_*^+, v_*^+) \leq_2 (u_+, v_+)$$

such that $\omega(u_*^-, v_*^-, f_1, f_2)$, $\omega(u_*^+, v_*^+, f_1, f_2)$ *are minimal and almost* 1*-covers of* $H(f_1, f_2)$ *(hence almost automorphic), and* $\mathcal{M}(\tilde{u}_*, \tilde{v}_*) \subset \mathcal{M}(a_1, b_1, c_1, a_2, b_2, c_2)$ *holds if* $(\bar{u}_*, \bar{v}_*, g_1, g_2) \in \omega(u_*^\pm, v_*^\pm, f_1, f_2)$ *is such that* $(\tilde{u}_*(t, x), \tilde{v}_*(t, x)) = (u(t, x; \bar{u}_*, \bar{v}_*, g_1, g_2), v(t, x; \bar{u}_*, \bar{v}_*, g_1, g_2))$ *is almost automorphic in* t. *In addition, if* a_i, b_i, c_i *are spatially homogeneous, then* $(u_*^-, v_*^-) = (u_*^+, v_*^+) = (u_*, v_*)$ *and* $(u(t; u_*, v_*, f_1, f_2), v(t; u_*, v_*, f_1, f_2))$ *is a globally stable positive almost periodic solution of* (4.1).

(2) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$ *and* $a_{2M} \leq \frac{a_{1L}b_{2L}}{b_{1M}}$, *then every positive solution of* (4.1) *converges to* $(u_{f_1}(t, x), 0)$.

(3) *If* $a_{1M} \leq \frac{c_{1L}a_{2L}}{c_{2M}}$ *and* $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, *then every positive solution of* (4.1) *converges to* $(0, v_{f_2}(t, x))$.

(4) *If* $a_1 = a_2$, $b_1 = b_2 = c_1 = c_2$, *and additionally* $k_1 = k_2$ *in the case where* a_i, b_i, c_i *are not spatially homogeneous, then there exists a stable continuous family of positive almost periodic solutions connecting* $(u_{f_1}(t, x), 0)$ *and* $(0, v_{f_2}(t, x))$.

The following results hold for the Dirichlet case.

THEOREM 5.2. *Consider* (4.1). *Suppose that* $Bu = u$, a_{iL}, b_{iL}, c_{iL} *are positive constants, and* $\lambda(k_1, a_1) > 0$, $\lambda(k_2, a_2) > 0$.

(1) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$, $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, $k_1 = k_2$, *and* $a_1 = a_2$ *(constant), then there are* (u_-, v_-), $(u_+, v_+) \in \mathrm{Int}X_+ \times \mathrm{Int}X_+$ *with* $(u_-, v_-) \ll_2 (u_+, v_+)$ *such that, for each* $(u_0, v_0) \in (X_+ \setminus \{0\}) \times X_+ \setminus \{0\}$ *and each* $(u, v, g_1, g_2) \in \omega(u_0, v_0, f_1, f_2)$,

$$(u_-, v_-) \leq_2 (u, v) \leq_2 (u_+, v_+)$$

(hence uniform persistence occurs). Moreover, there exist (u_*^-, v_*^-), (u_*^+, v_*^+) *with*

$$(u_-, v_-) \leq_2 (u_*^-, v_*^-) \leq_2 (u_*^+, v_*^+) \leq_2 (u_+, v_+)$$

such that $\omega(u_*^-, v_*^-, f_1, f_2)$, $\omega(u_*^+, v_*^+, f_1, f_2)$ *are minimal and almost* 1*-covers of* $H(f_1, f_2)$ *(hence almost automorphic), and* $\mathcal{M}(\tilde{u}_*, \tilde{v}_*) \subset \mathcal{M}(a_1, b_1, c_1, a_2, b_2, c_2)$ *if* $(\bar{u}_*, \bar{v}_*, g_1, g_2) \in \omega(u_*^\pm, v_*^\pm, f_1, f_2)$ *is such that* $(\tilde{u}_*(t, x), \tilde{v}_*(t, x)) = (u(t, x; \bar{u}_*, \bar{v}_*, g_1, g_2), v(t, x; \bar{u}_*, \bar{v}_*, g_1, g_2))$ *is almost automorphic in* t.

(2) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$, $a_{2M} \leq \frac{b_{2L}a_{1L}}{b_{1M}}$, $k_2 \geq k_1$, *and* $a_{1L} \geq a_2$, *then every positive solution converges to* $(u_{f_1}(t, x), 0)$.

(3) *If* $a_{1M} \leq \frac{c_{1L}a_{2L}}{c_{2M}}$, $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, $k_1 \geq k_2$, *and* $a_{2L} \geq a_1$, *then every positive solution converges to* $(0, v_{f_2}(t, x))$.

(4) *If* $k_1 = k_2$, $a_1 = a_2$, *and* $b_1 = b_2 = c_1 = c_2$, *then there exists a stable continuous family of positive almost periodic solutions connecting* $(u_{f_1}(t, x), 0)$ *and* $(0, v_{f_2}(t, x))$.

Proof of Theorem 5.1. (1) First, by $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$ and $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, we have

$$\frac{c_{1L}}{c_{2M}} \le \frac{c_{1M}}{c_{2L}} < \frac{a_{1L}}{a_{2M}} \le \frac{a_{1M}}{a_{2L}} < \frac{b_{1L}}{b_{2M}} \le \frac{b_{1M}}{b_{2L}}.$$

Hence, by Lemma 4.7, $(u^{\pm}(t;u_0,v_0), v^{\pm}(t;u_0,v_0))$ converges to a unique strictly positive equilibrium $(u_{\pm}, v_{\pm})$ of $(4.8)_{\pm}$ for each $(u_0,v_0) \in (X_+ \setminus \{0\}) \times (X_+ \setminus \{0\})$.

Next, Lemma 4.5 shows

$$(u^-(t,\cdot;u_0,v_0), v^-(t,\cdot;u_0,v_0)) \le_2 (u(t,\cdot;u_0,v_0,f_1,f_2), v(t,\cdot;u_0,v_0,f_1,f_2))$$

and

$$(u(t,\cdot;u_0,v_0,f_1,f_2), v(t,\cdot;u_0,v_0,f_1,f_2)) \le_2 (u^+(t,\cdot;u_0,v_0), v^+(t,\cdot;u_0,v_0))$$

for $t \ge 0$ and all $(u_0,v_0) \in (X_+ \setminus \{0\}) \times (X_+ \setminus \{0\})$, hence

$$(u_-, v_-) \le_2 (u,v) \le_2 (u_+, v_+)$$

for $(u,v,g_1,g_2) \in \omega(u_0,v_0,f_1,f_2)$.

Now we have $(u_-,v_-) \le_2 (u,v)$ for every $(u,v,g_1,g_2) \in \omega(u_-,v_-,f_1,f_2)$. By Lemma 2.2(3), $\omega(u_-,v_-,\ f_1,f_2)$ is an almost 1-cover of $H(f_1,f_2)$. Similarly, we obtain that $\omega(u_+,v_+,f_1,f_2)$ is an almost 1-cover of $H(f_1,f_2)$. Therefore, there are $(u_*^{\pm},,v_*^{\pm},f_1,f_2) \in \omega(u_{\pm},v_{\pm},f_1,f_2)$ such that

$$(u_-,v_-) \le_2 (u_*^-,v_*^-) \le_2 (u_*^+,v_*^+) \le_2 (u_+,v_+)$$

and $\omega(u_*^{\pm},v_*^{\pm},f_1,f_2)$ are minimal and almost 1-covers of $H(f_1,f_2)$.

Note that if $(\bar u_*,\bar v_*,g_1,g_2) \in \omega(u_*^{\pm},v_*^{\pm},f_1,f_2)$ is such that

$$\{(\bar u_*,\bar v_*,g_1,g_2)\} = \omega(u_*^{\pm},v_*^{\pm},f_1,f_2) \cap (X_+ \times X_+ \times \{(g_1,g_2)\}), \tag{5.1}$$

then, by the definition of almost automorphic functions (see section 2.1),

$$(\tilde u_*(t,x), \tilde v_*(t,x)) = (u(t,x;\bar u_*,\bar v_*,g_1,g_2), v(t,x;\bar u_*,\bar v_*,g_1,g_2)) \tag{5.2}$$

is uniformly almost automorphic in t (hence $\omega(u_*^{\pm},v_*^{\pm},f_1,f_2)$ is almost automorphic). Moreover, by Lemma 2.1, $\mathcal{M}(\tilde u_*,\tilde v_*) \subset \mathcal{M}(a_1,b_1,c_1,a_2,b_2,c_2)$. Conversely, if $(\bar u_*,\bar v_*,g_1,g_2) \in \omega(u_*^{\pm},v_*^{\pm},f_1,f_2)$ is such that $(\tilde u_*(t,x),\ \tilde v_*(t,x))$ in (5.2) is uniformly almost automorphic in t, then (5.1) must hold, and hence one has $\mathcal{M}(\tilde u_*,\tilde v_*) \subset \mathcal{M}(a_1,b_1,c_1,a_2,b_2,c_2)$. Otherwise, there is $(\bar u_*^1,\bar v_*^1,g_1,g_2) \in \omega(u_*^{\pm},v_*^{\pm},f_1,f_2)$ with $(\bar u_*^1,\bar v_*^1) \ne (\bar u_*,\bar v_*)$. Let $(g_1^0,g_2^0) \in H(f_1,f_2)$ be such that

$$\omega(u_*^{\pm},v_*^{\pm},f_1,f_2) \cap (X_+ \times X_+ \times \{(g_1^0,g_2^0)\}) = \{(\bar u_*^0,\bar v_*^0,g_1^0,g_2^0)\}$$

is a singleton. Let $\beta_n \to \infty$ be such that

$$\Pi_{\beta_n}(\bar u_*^0,\bar v_*^0,g_1^0,g_2^0) \to (\bar u_*^1,\bar v_*^1,g_1,g_2)$$

and $\alpha_n = -\beta_n$. Then, given any subsequence $\{\alpha_{n_k}\} \subset \{\alpha_n\}$,

$$\lim_{m\to\infty}\lim_{k\to\infty} \Pi_{-\alpha_{n_m}}\Pi_{\alpha_{n_k}}(\bar u_*,\bar v_*,g_1,g_2) = (\bar u_*^1,\bar v_*^1,g_1,g_2) \ne (\bar u_*,\bar v_*,g_1,g_2),$$

which contradicts the almost automorphy of $(\tilde u^*(t,x),\tilde v^*(t,x))$ (see section 2.1).

Finally, if a_i, b_i and c_i $(i = 1, 2)$ are spatially homogeneous, by [22], $(u_*^-, v_*^-) = (u_*^+, v_*^+) = (u_*, v_*)$ and $(u(t; u_*, v_*, f_1, f_2), v(t; u_*, v_*, f_1, f_2))$ is a globally stable positive almost periodic solution of (4.1).

(2) First, $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$ and $a_{2M} \le \frac{a_{1L}b_{2L}}{b_{1M}}$ imply

$$\frac{c_{1M}}{c_{2L}} < \frac{a_{1L}}{a_{2M}} \quad \text{and} \quad \frac{a_{1L}}{a_{2M}} \ge \frac{b_{1M}}{b_{2L}}.$$

Therefore, by Lemma 4.7, every positive solution of $(4.8)_-$ converges to $(u_-^*, 0) \equiv (\frac{a_{1L}}{b_{1M}}, 0)$.

Next, let $(u_0, v_0) \in (X_+ \setminus \{0\}) \times (X_+ \setminus \{0\})$. Lemma 4.5 yields

$$(u_-^*, 0) \le_2 (u, v) \le_2 (u_{g_1}, 0)$$

for every $(u, v, g_1, g_2) \in \omega(u_0, v_0, f_1, f_2)$. Assume that $(u_0, v_0) \le_2 (u_-^*, 0)$. Then, by Lemma 2.2(3), $\omega(u_0, v_0, f_1, f_2)$ is an almost 1-cover of $H(f_1, f_2)$. By the arguments of Lemma 3.7, $\omega(u_0, v_0, f_1, f_2)$ is a 1-cover of $H(f_1, f_2)$ and hence is minimal. By Lemma 4.3, we must have $\omega(u_0, v_0, f_1, f_2) = E_1$, and hence $(u(t, x; u_0, v_0, f_1, f_2), v(t, x; u_0, v_0, f_1, f_2))$ converges to $(u_{f_1}(t, x), 0)$.

Finally, for any $(u_0, v_0) \in \mathrm{Int}X_+ \times \mathrm{Int}X_+$, there is $(\tilde{u}_0, \tilde{v}_0) \in \mathrm{Int}X_+ \times \mathrm{Int}X_+$ such that $(\tilde{u}_0, \tilde{v}_0) \ll_2 (u_0, v_0)$ and $(\tilde{u}_0, \tilde{v}_0) \le_2 (u_-^*, 0)$. By the above arguments, $\omega(\tilde{u}_0, \tilde{v}_0, f_1, f_2) = E_1$, and we must have $\omega(u_0, v_0, f_1, f_2) = E_1$, and hence every positive solution of (4.1) converges to $(u_{f_1}(t, x), 0)$.

(3) The proof can be derived by similar arguments as in (2).

(4) If a_i, b_i, c_i are spatially homogeneous, the proof follows from [22]. Otherwise, note that $w = u + v$ satisfies

$$\begin{cases} w_t = k\Delta w + w(a - bw), & x \in \Omega, \\ \dfrac{\partial w}{\partial n} = 0, & x \in \partial\Omega, \end{cases}$$

where $k = k_1 = k_2$, $a = a_1 = a_2$, $b = b_1 = b_2 = c_1 = c_2$. (4) then follows from Theorem 3.3. □

Proof of Theorem 5.2. (1) Since $a_1 = a_2$ are constant, $\lambda(k_i, a_{iL(M)}) > 0$ for $i = 1, 2$. Then, by Lemma 4.8, $(u^{\pm}(t, x; u_0, v_0), v^{\pm}(t, x; u_0, v_0))$ converges to a unique strictly positive equilibrium $(u_{\pm}, v_{\pm})$ of $(4.8)_{\pm}$ for every $(u_0, v_0) \in (X_+ \setminus \{0\}) \times (X_+ \setminus \{0\})$. The rest of the proof now follows by employing the same arguments as in the proof of Theorem 5.1(1).

(2) By $k_2 \ge k_1$, $a_{1L} \ge a_2$ and Lemma 2.3, we have $\lambda(k_1, a_{1L}) \ge \lambda(k_2, a_{1L}) \ge \lambda(k_2, a_2) > 0$ and $\lambda(k_2, a_{2M}) \ge \lambda(k_2, a_2) > 0$. By Lemma 4.8, for each $(u_0, v_0) \in (X_+ \setminus \{0\}) \times (X_+ \setminus \{0\})$, $(u^-(t, x; u_0, v_0), v^-(t, x; u_0, v_0))$ converges to $(u_-^*, 0)$, where u_-^* is the unique positive equilibrium of (3.1) with $k = k_1$, $a = a_{1L}$, $b = b_{1M}$, and $Bu = u$. The rest of the proof then follows from the same arguments as in Theorem 5.1(2).

(3) The proof can be derived by similar arguments as in the proof of (2).

(4) The proof can be derived by similar arguments as in the proof of Theorem 5.1(4). □

6. Single species population and two species competition models with recurrent time dependence. In this section, we state results similar to those of Theorems 3.3, 5.1, and 5.2 for more general time dependent single species population and two species competition models. They can be derived by the approach we have

developed in the previous sections. We deal with the case where the reaction terms in (1.1) and (1.3) exhibit merely recurrent time dependence.

Throughout this section, we assume that a_i, b_i, c_i $(i = 1, 2)$ in (1.1) and a, b in (1.3) are recurrent in t, $b(t, x) \geq \delta$ for some $\delta > 0$. Let $a_{iL(M)}$, $b_{iL(M)}$, and $c_{iL(M)}$ be as in $(1.4)_1$–$(1.4)_3$.

First, by arguments similar to those in the proofs of Theorem 3.3 and Corollary 3.4, one obtains the following result.

THEOREM 6.1. *Consider* (1.3). *One and only one of the following alternatives occurs.*

(1) *Every positive solution converges to a unique strictly positive recurrent solution* $u^*(t, x)$ *whose hull is a* 1*-cover of the hull of* (a, b).

(2) *Every positive solution converges to the trivial solution* $u = 0$.

(3) *Every positive solution is neither bounded away from the trivial solution nor converges to it.*

Next, consider (1.1) with $Bu = \frac{\partial u}{\partial n}$. If $a_{1L} > 0$ and $a_{2L} > 0$, then by Theorem 6.1 and the comparison principle for parabolic equations, (1.3) with $a = a_1$ and $b = b_1$ ($a = a_2$ and $b = c_2$) has a unique strictly positive recurrent solution $u^*_N(t, x)$ ($v^*_N(t, x)$). Following arguments similar to those in proving Theorem 5.1, we have the following result.

THEOREM 6.2. *Consider* (1.1). *Assume that* $Bu = \frac{\partial u}{\partial n}$ *and* a_{iL}, b_{iL}, $c_{iL} > 0$ $(i = 1, 2)$.

(1) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$ *and* $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, *then uniform persistence occurs. Moreover, there is a strictly positive recurrent solution* $(u_*(t, x), v_*(t, x))$ *whose hull is an almost* 1*-cover of the hull of* $(a_1, b_1, c_1, a_2, b_2, c_2)$.

(2) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$ *and* $a_{2M} \leq \frac{a_{1L}b_{2L}}{b_{1M}}$, *then every positive solution converges to* $(u^*_N(t, x), 0)$.

(3) *If* $a_{1M} \leq \frac{c_{1L}a_{2L}}{c_{2M}}$ *and* $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, *then every positive solution converges to* $(0, v^*_N(t, x))$.

Finally, consider (1.1) with $Bu = u$. Assume that $\lambda(k_1, H(a_1)) > 0$ and $\lambda(k_1, H(a_2)) > 0$. If $k_2 \geq k_1$ and $a_{1L} \geq a_2$ ($k_1 \geq k_2$ and $a_{2L} \geq a_1$), then $\lambda(k_1, a_{1L}) \geq \lambda(k_2, a_{1L}) \geq \lambda(k_2, H(a_2)) > 0$ ($\lambda(k_2, a_{2L}) \geq \lambda(k_1, a_{2L}) \geq \lambda(k_1, H(a_1)) > 0$). Again, by Theorem 6.1 and the comparison principle for parabolic equations, (1.3) with $a = a_1$ and $b = b_1$ ($a = a_2$ and $b = c_2$) has a unique strictly positive recurrent solution $u^*_D(t, x)$ ($v^*_D(t, x)$). Arguments similar to those in proving Theorem 5.2 yield the following result.

THEOREM 6.3. *Consider* (1.1) *with* a_i, b_i, c_i $(i = 1, 2)$ *being recurrent. Assume that* $Bu = u$, a_{iL}, b_{iL}, $c_{iL} > 0$ $(i = 1, 2)$, *and* $\lambda(k_1, H(a_1)) > 0$, $\lambda(k_2, H(a_2)) > 0$.

(1) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$, $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, $k_1 = k_1$, *and* $a_1 = a_2$ *(constant), then uniform persistence occurs. Moreover, there is a strictly positive recurrent solution* $(u_*(t, x), v_*(t, x))$ *whose hull is an almost* 1*-cover of the hull of* $(a_1, b_1, c_1, a_2, b_2, c_2)$.

(2) *If* $a_{1L} > \frac{c_{1M}a_{2M}}{c_{2L}}$, $a_{2M} \leq \frac{a_{1L}b_{2L}}{b_{1M}}$, $k_2 \geq k_1$, *and* $a_{1L} \geq a_2$, *then every positive solution converges to* $(u^*_D(t, x), 0)$.

(3) *If* $a_{1M} \leq \frac{c_{1L}a_{2L}}{c_{2M}}$, $a_{2L} > \frac{a_{1M}b_{2M}}{b_{1L}}$, $k_1 \geq k_2$, *and* $a_{2L} \geq a_1$, *then every positive solution converges to* $(0, v^*_D(t, x))$.

REFERENCES

[1] S. AHMAD, *Convergence and ultimate bounds of solutions of the nonautonomous Volterra-Lotka competition equations*, J. Math. Anal. Appl., 127 (1987), pp. 377–387.

[2] S. AHMAD, *On the nonautonomous Volterra-Lotka competition equations*, Proc. Amer. Math. Soc., 117 (1993), pp. 199–204.

[3] S. Ahmad and A. Lazer, *Asymptotic behavior of solutions of periodic competition diffusion system*, Nonlinear Anal., 13 (1989), pp. 263–284.
[4] C. Alvarez and A. Lazer, *An application of topological degree to the periodic competing species problem*, J. Austral. Math. Soc. Ser. B, 28 (1986), pp. 202–219.
[5] P. W. Bates, K. Lu, and C. Zeng, *Existence and persistence of invariant manifolds for semiflows in Banach space*, Mem. Amer. Math. Soc., 135 (1998).
[6] R. S. Cantrell, C. Cosner and V. Hutson, *Permanence in ecological systems with spatial heterogeneity*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 533–559.
[7] S.-N. Chow and K. Lu, *Invariant manifolds for flows in Banach spaces*, J. Differential Equations, 74 (1988), pp. 285–317.
[8] C. Cosner and A. C. Lazer, *Stable coexistence states in the Volterra–Lotka competition model with diffusion*, SIAM J. Appl. Math., 44 (1984), pp. 1112–1132.
[9] H. Engler and G. Hetzer, *Convergence to equilibria for a class of reaction-diffusion systems*, Osaka J. Math., 29 (1992), pp. 471–481.
[10] G. Fan and A. Leung, *Existence and stability of periodic solutions for competing species diffusion systems with Dirichlet boundary conditions*, Appl. Anal., 39 (1990), pp. 119–149.
[11] A. M. Fink, *Almost Periodic Differential Equations*, Lecture Notes in Math. 377, Springer-Verlag, Berlin, Heidelberg, New York, 1974.
[12] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[13] K. Gopalsamy, *Global asymptotic stability in a periodic Lotka-Volterra system*, J. Austral. Math. Soc. Ser. B, 27 (1985), pp. 66–72.
[14] K. Gopalsamy, *Global asymptotic stability in an almost-periodic Lotka-Volterra system*, J. Austral. Math. Soc. Ser. B, 27 (1986), pp. 346–360.
[15] K. Gopalsamy and X. Z. He, *Oscillations and convergence in an almost periodic competition systems*, Acta Appl. Math., 46 (1977), pp. 247–266.
[16] J. K. Hale and P. Waltman, *Persistence in infinite-dimensional systems*, SIAM J. Math. Anal., 20 (1989), pp. 388–395.
[17] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.
[18] P. Hess, *Periodic-Parabolic Boundary Value Problems and Positivity*, Pitman Res. Notes Math. Ser. 247, Longman Scientific and Technical, Harlow, UK, 1991.
[19] P. Hess, *Asymptotics in semilinear periodic diffusion equations with Dirichlet or Robin boundary conditions*, Arch. Ration. Mech. Anal., 116 (1991), pp. 91–99.
[20] P. Hess and A. Lazer, *On an abstract competition model and applications*, Nonlinear Anal., 16 (1991), pp. 917–940.
[21] P. Hess and H. Weinberger, *Convergence to spatial-temporal clines in the Fisher equation with time-periodic fitnesses*, J. Math. Biol., 28 (1990), pp. 83–98.
[22] G. Hetzer and W. Shen, *Convergence in almost periodic competition diffusion systems*, J. Math. Anal. Appl., 262 (2001), pp. 307–338.
[23] G. Hetzer, W. Shen, and S. Zhu, *Convergence in random and stochastic parabolic equations of Fisher and Kolmogorov types*, J. Dynam. Differential Equations, 14 (2002), pp. 139–188.
[24] M. W. Hirsch, *Stability and convergence in strongly monotone dynamical systems*, J. Reine Angew. Math., 383 (1988), pp. 1–58.
[25] S. B. Hsu, H. L. Smith, and P. Waltman, *Competitive exclusion and coexistence for competitive systems on ordered Banach spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 4083–4094.
[26] V. Hutson, J. López-Gómez, K. Mischaikow, and G. Vickers, *Limit behaviour for a competing species problem with diffusion*, in Dynamical Systems and Applications, World Sci. Ser. Appl. Anal. 4, World Sci. Publishing, River Edge, NJ, 1995, pp. 343–358.
[27] V. Hutson and K. Schmitt, *Permanence and the dynamics of biological systems*, Math. Biosci., 111 (1992), pp. 1–71.
[28] V. Hutson, W. Shen, and G. T. Vickers, *Estimates for the principal spectrum point for certain time-dependent parabolic operators*, Proc. Amer. Math. Soc., 129 (2000), pp. 1669–1679.
[29] H. Matano, *Strong comparison principle in nonlinear parabolic equations*, in Nonlinear Parabolic Equations: Qualitative Properties of Solutions, L. Boccardo and A. Tesei, eds., Longman Scientific and Technical, Harlow, UK, 1987, pp. 148–155.
[30] P. De Mottoni and A. Schiaffino, *Competition systems with periodic coefficients: A geometric approach*, J. Math. Biol., 11 (1981), pp. 319–335.
[31] C. V. Pao, *Coexistence and stability of a competition-diffusion system in population dynamics*, J. Math. Anal. Appl., 83 (1981), pp. 54–76.

[32] P. Poláčik and I. Tereščák, *Exponential separation and invariant bundles for maps in ordered Banach spaces with applications to parabolic equations*, J. Dynam. Differential Equations, 5 (1993), pp. 279–303.
[33] P. Poláčik and I. Tereščák, *Convergence to cycles as a typical asymptotic behavior in smooth strongly monotone discrete-time dynamical systems*, Arch. Ration. Mech. Anal., 116 (1991), pp. 339–360.
[34] M. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
[35] G. Sell, *Nonautonomous differential equations and topological dynamics*, I, Trans. Amer. Math. Soc., 127 (1967), pp. 241–262.
[36] G. Sell, *Nonautonomous differential equations and topological dynamics*, II, Trans. Amer. Math. Soc., 127 (1967), pp. 263–283.
[37] W. Shen and Y. Yi, *Convergence in almost periodic Fisher and Kolmogorov models*, J. Math. Biol., 37 (1998), pp. 84–102.
[38] W. Shen and Y. Yi, *Almost automorphic and almost periodic dynamics in skew-product semiflows*, Mem. Amer. Math. Soc., 136 (1998), pp. 23–52.
[39] W. Shen and Y. Yi, *On minimal sets of scalar parabolic equations with skew-product structures*, Trans. Amer. Math. Soc., 347 (1995), pp. 4413–4431.
[40] H. L. Smith and H. R. Thieme, *Quasi convergence and stability for order-preserving semiflows*, SIAM J. Math. Anal., 21 (1990), pp. 673–692.
[41] H. L. Smith and P. Waltman, *The Theory of the Chemostat*, Cambridge Stud. Math. Biol. 13, Cambridge University Press, Cambridge, UK, 1995.
[42] P. Takáč, *Discrete monotone dynamics and time-periodic competition between two species*, Differential Integral Equations, 10 (1997), pp. 547–576.
[43] P. Takáč, *private communication*, Auburn University, 1998.
[44] W. A. Veech, *Almost automorphic functions on groups*, Amer. J. Math., 87 (1965), pp. 719–751.
[45] Y. Yi, *Almost automorphic and almost periodic dynamics in skew-product semiflows*, Mem. Amer. Math. Soc., 136 (1998), pp. 1–22.
[46] X.-Q. Zhao, *Uniform persistence and periodic coexistence states in infinite-dimensional periodic semiflows with applications*, Canad. Appl. Math. Quart., 3 (1995), pp. 473–495.
[47] X.-Q. Zhao and V. Hutson, *Permanence in Kolmogorov periodic predator-prey models with diffusion*, Nonlinear Anal., 23 (1994), pp. 651–668.
[48] L. Zhou and C. V. Pao, *Asymptotic behavior of a competition-diffusion system in population dynamics*, Nonlinear Anal., 6 (1982), pp. 1163–1184.

SIAM J. MATH. ANAL.
Vol. 34, No. 1, pp. 228–238

TRIPLE VARIATIONAL PRINCIPLES FOR EIGENVALUES OF SELF-ADJOINT OPERATORS AND OPERATOR FUNCTIONS*

DAVID ESCHWÉ† AND HEINZ LANGER†

Abstract. We derive triple variational principles for the eigenvalues of a self-adjoint operator pencil, which also allow a characterization of discrete eigenvalues within a gap of the essential spectrum. In the general case, we can prove only an inequality; the equality sign is shown to hold in four particular situations.

Key words. eigenvalue, variational principle, operator pencil

AMS subject classifications. 47A75, 49R50, 47A56

PII. S0036141001387744

1. Introduction. It is well known that the discrete eigenvalues of a self-adjoint operator A on some Hilbert space $\mathcal{H}$, which lie below or above the essential spectrum of A, can be characterized by double variational principles applied to the Rayleigh quotients $\frac{(Ax,x)}{(x,x)}$, $x \in \mathcal{H}$, $x \neq 0$. For example, if the eigenvalues below the minimum of the essential spectrum of A are denoted by

$$\lambda_1 \leq \lambda_2 \leq \cdots, \tag{1.1}$$

counted according to their multiplicities, then

$$\lambda_j = \sup_{\substack{\mathcal{V}\subset\mathcal{H}\\ \dim\mathcal{V}=j-1}} \inf_{\substack{x\neq 0\\ x\perp\mathcal{V}}} \frac{(Ax,x)}{(x,x)}, \qquad j = 1,2,\ldots.$$

These formulas have been generalized for sufficiently smooth self-adjoint operator functions L, defined on some interval Δ of the real axis, under the assumption that for some $\alpha \in \Delta$ the operator $L(\alpha)$ is uniformly positive or uniformly negative (see, e.g., [M], [BEL]): in this case the discrete eigenvalues, numbered according to their multiplicity and their distance from α, can be characterized by such double variational principles applied to the zeros of the scalar functions $(L(\,\cdot\,)x,x)$, $x \in \mathcal{H}$, $x \neq 0$. They have been generalized further to the case where $L(\alpha)$ is not definite but has a finite number of positive or of negative eigenvalues (see [BEL]). In the latter case, in the formulas an index shift, corresponding to the number of these eigenvalues, appears.

On the other hand, already in 1970 Phillips [P] and, subsequently, Textorius [T] proved a *triple* variational principle for eigenvalues of positive compact operators on Krein spaces. For example, given a positive self-adjoint operator A on a Krein space $(\mathcal{K},[\,\cdot\,,\,\cdot\,])$ such that below the essential spectrum of A there are isolated eigenvalues as in (1.1), then

$$\lambda_j = \sup_{\mathcal{M}\in\mathbf{M}^{--}} \sup_{\substack{\mathcal{V}\subset\mathcal{M}\\ \dim\mathcal{V}=j-1}} \inf_{\substack{x\neq 0\\ x\perp\mathcal{V}}} \frac{[Ax,x]}{[x,x]}, \qquad j = 1,2,\ldots;$$

*Received by the editors April 9, 2001; accepted for publication (in revised form) April 16, 2002; published electronically September 24, 2002. This work was supported by the Research Training Network HPRN-CT-2000-00116 of the European Union.

http://www.siam.org/journals/sima/34-1/38774.html

†Vienna University of Technology, Institute of Analysis and Technical Mathematics,Wiedener Hauptstrasse 8-10, A-1040 Wien, Austria (deschwe@gmx.at, hlanger@mail.zserv.tuwien.ac.at).

here $\mathbf{M}^{--}$ denotes the set of all maximal negative subspaces of $\mathcal{K}$. This result can be understood as corresponding to a situation where the function L is linear in the parameter λ and for a certain α the operator $L(\alpha)$ has infinitely many positive and infinitely many negative spectral points. On the other hand, it is not hard to see that the above mentioned double variational principles with index shift can also be replaced by triple variational principles.

In the present paper we show that such triple variational principles do also hold in other situations for discrete eigenvalues which lie in a gap of the essential spectrum. In section 2 we prove a general triple variational inequality for discrete eigenvalues of a continuous self-adjoint operator function L which satisfies Assumptions 1–3 (listed in section 2 below). These discrete eigenvalues are numbered starting from a certain point $\alpha \in \rho(L)$, and in the variational principle the maximal $L(\alpha)$-nonnegative subspaces play an important role. We do not know if in the general situation of section 2 such a maximal $L(\alpha)$-nonnegative subspace exists for which the inequalities become equalities. This we can show only in particular situations in section 3. Namely, we can show it for the discrete eigenvalues in a gap of the essential spectrum of a self-adjoint operator, for a nonnegative operator in a Krein space (this situation corresponds to the results of Phillips and Textorius mentioned above), for a special class of quadratic operator pencils, and for a self-adjoint block operator matrix

$$\widetilde{A} = \begin{pmatrix} A & B \\ B^* & D \end{pmatrix}.$$

In the latter case, double variational principles for eigenvalues in a certain gap of the essential spectrum of $\widetilde{A}$ were proved by Griesemer and Siedentop [GS] if the numerical ranges of A and D overlap in at most one point. Here we show that triple variational principles also allow us to characterize certain discrete eigenvalues of $\widetilde{A}$ if the numerical ranges of A and D overlap in an interval.

In this paper we restrict ourselves to bounded operators. In a subsequent publication pencils of unbounded self-adjoint operators will be considered. They include Hain–Lüst-type equations for partial or ordinary differential operators; e.g.,

$$-y'' + \lambda y - \frac{qy}{u - \lambda} = 0 \quad \text{on} \quad [0,1], \qquad y(0) = y(1) = 0,$$

with real continuous functions q, u (see [ALM]). Here the essential spectrum consists of the range of the function u, and the eigenvalues $\lambda_1 \le \lambda_2 \le \cdots$ to the right of this essential spectrum should be characterized by triple variational principles from the left. They also include quadratic pencils arising in the consideration of beams with inner and outer damping (so-called Vogt material), e.g., from the equation

$$\alpha \frac{\partial^5 u}{\partial t \partial x^4} + \frac{\partial^4 u}{\partial x^4} + \frac{\partial}{\partial x} g(x) \frac{\partial u}{\partial x} + k(x) \frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial t^2} = 0$$

with appropriate boundary and initial conditions (see [Pi2]). Finally, these principles can also be applied to the problem considered in [LM].

2. A general inequality. Let $\mathcal{H}$ be a Hilbert space. We make the following assumptions.

Assumption 1. The operator function L is defined and continuous in the operator norm on the interval $[\alpha, \beta)$, and its values are self-adjoint operators on $\mathcal{H}$ and $0 \in \rho(L(\alpha))$.

It follows that for some $\alpha' > \alpha$ the interval $[\alpha, \alpha')$ belongs to $\rho(L)$. We allow $\beta = \infty$. The point $\lambda \in \mathcal{C}$ is an *eigenvalue* of the operator function L if 0 is an eigenvalue of $L(\lambda)$, and a *normal eigenvalue* of L if 0 is a normal eigenvalue of $L(\lambda)$; recall (see [GK]) that the latter means that the algebraic eigenspace $\mathcal{L}$ of $L(\lambda)$ at 0 is finite-dimensional and that the space $\mathcal{H}$ is the direct sum of $\mathcal{L}$ and an invariant subspace $\mathcal{N}$ of $L(\lambda)$ such that $0 \in \rho(L(\lambda)|\mathcal{N})$. Further, the point $\lambda \in \mathcal{C}$ belongs to the *essential spectrum* $\sigma_{ess}(L)$ of the operator function L if $0 \in \sigma_{ess}(L(\lambda))$. If the essential spectrum $\sigma_{ess}(L)$ in $[\alpha, \beta)$ is not empty we set $\lambda_e := \min \sigma_{ess}(L) \cap [\alpha, \beta)$; otherwise, if there is no essential spectrum of L in $[\alpha, \beta)$, then $\lambda_e := \beta$. By definition of the essential spectrum of an operator function and Assumption 1, in the interval (α, λ_e) the spectrum of L is discrete; that is, it consists of normal eigenvalues of $L(\lambda)$, and λ_e is their only possible accumulation point.

If we equip the space $\mathcal{H}$ with the inner product $[\,\cdot\,,\,\cdot\,]_\alpha := (L(\alpha)\cdot\,,\,\cdot\,)$, by Assumption 1 it becomes a Krein space (see [B], [AI]) which we denote by $\mathcal{K}_\alpha$. A natural canonical decomposition of this Krein space is given by $\mathcal{K}_\alpha = \mathcal{H}_+ \oplus \mathcal{H}_-$, where $\mathcal{H}_+$ is the spectral invariant subspace of $L(\alpha)$ corresponding to $(0, +\infty)$ and $\mathcal{H}_-$ is the spectral invariant subspace of $L(\alpha)$ corresponding to $(-\infty, 0)$. In this section, the set of all maximal nonnegative subspaces of this Krein space $\mathcal{K}_\alpha$, which are also called *maximal $L(\alpha)$-nonnegative subspaces* of $\mathcal{H}$, is denoted by $\mathbf{M}_\alpha^+$.

A function φ, considered on a real interval, is said to be *decreasing at value zero* if $\varphi(\lambda_0) = 0$ implies that $\varphi(\lambda) > 0$ if $\lambda < \lambda_0$ and $\varphi(\lambda) < 0$ if $\lambda > \lambda_0$.

Assumption 2. For each $x \in \mathcal{H}$, $x \neq 0$, the function $\varphi_x : \varphi_x(\lambda) := (L(\lambda)x, x)$, $\lambda \in [\alpha, \beta)$, is decreasing at value zero.

It follows that each function φ_x, $x \neq 0$, has at most one zero in the interval $[\alpha, \beta)$; this zero is denoted by $p(x)$. If $\varphi_x(\alpha) > 0$ and the function φ_x does not have a zero in $[\alpha, \beta)$ we put $p(x) = +\infty$. Evidently, if $x \in \mathcal{H}$, $x \neq 0$, and $\gamma \neq 0$ is a complex number, then $p(\gamma x) = p(x)$. Also, the convention $\min \emptyset = +\infty$ is used.

Sometimes we need the following assumption.

Assumption 3. If $\lambda_0 \in (\alpha, \beta)$ is fixed, for each $\varepsilon > 0$ such that $(\lambda_0 - \varepsilon, \lambda_0 + \varepsilon) \subset (\alpha, \beta)$ there exists a $\delta(\varepsilon) > 0$ such that $\|x\| = 1$, $|\varphi_x(\lambda_0)| \leq \delta(\varepsilon)$ implies that φ_x has a zero in the interval $(\lambda_0 - \varepsilon, \lambda_0 + \varepsilon)$.

For any subspace $\mathcal{M}$ of $\mathcal{H}$, by $\mathcal{M}^1$ we denote the unit sphere of $\mathcal{M}$ that is the set of all elements $x \in \mathcal{M}$ with $\|x\| = 1$.

THEOREM 2.1. *Under Assumptions* 1–3, *for each subspace* $\mathcal{M} \in \mathbf{M}_\alpha^+$ *we have*

$$\inf_{x \in \mathcal{M}^1} p(x) \leq \min \sigma(L) \cap [\alpha, \beta).$$

Proof. Denote $a := \min \sigma(L) \cap [\alpha, \beta)$ and assume to the contrary that

$$\inf_{x \in \mathcal{M}^1} p(x) > a. \tag{2.1}$$

First we suppose that a is an eigenvalue of L: $L(a)x_0 = 0$, $\|x_0\| = 1$. Then $p(x_0) = a$ and, by (2.1), $x_0 \notin \mathcal{M}$. We consider $\mathcal{M}' := \operatorname{span}\{\mathcal{M}, x_0\}$. Since $L(a)$ is self-adjoint and $L(a)x_0 = 0$, for an arbitrary element $x \in \mathcal{M}$ we obtain

$$\begin{aligned}(L(a)(x + x_0), x + x_0) &= (L(a)x, x) + (L(a)x_0, x) + (L(a)x, x_0) + (L(a)x_0, x_0)\\ &= (L(a)x, x) \geq 0.\end{aligned}$$

It follows that $(L(\alpha)(x + x_0), x + x_0) \geq 0$, hence $\mathcal{M}'$ is an $L(\alpha)$-nonnegative subspace, a contradiction to the fact that $\mathcal{M}$ is a maximal $L(\alpha)$-nonnegative subspace.

In the general case, for the point a there exists a sequence (y_n) in $\mathcal{H}$, $\|y_n\| = 1$, such that $\|L(a)y_n\| \to 0$ if $n \to \infty$. According to Assumption 3 and (2.1) there exists a $c > 0$ such that $\varphi_x(a) \geq c\|x\|^2$ for all $x \in \mathcal{M}$. Indeed, otherwise in $\mathcal{M}$ there would exist a sequence of elements x_n, $\|x_n\| = 1$, such that $\varphi_{x_n}(a) \downarrow 0$, and Assumption 3 would imply that $p(x_n) \to a$, which is impossible because of (2.1). Now we obtain for $x \in \mathcal{M}$

$$\begin{aligned}(L(a)(x+y_n), x+y_n) &= (L(a)x, x) + (L(a)x, y_n) + (L(a)y_n, x) + (L(a)y_n, y_n)\\ &\geq c\|x\|^2 - 2\|x\|\,\|L(a)y_n\| + (L(a)y_n, y_n)\\ &\geq \left(\sqrt{c}\|x\| - \frac{\|L(a)y_n\|}{\sqrt{c}}\right)^2 - \frac{\|L(a)y_n\|^2}{c} - \|L(a)y_n\|\\ &\geq -\left(\frac{\|L(a)y_n\|^2}{c} + \|L(a)y_n\|\right).\end{aligned}$$

If we choose $\varepsilon > 0$ such that $(a-\varepsilon, a+\varepsilon) \subset [\alpha, \beta)$, $\delta(\varepsilon)$ according to Assumption 3, and, finally, n such that $\frac{\|L(a)y_n\|^2}{c} + \|L(a)y_n\| < \delta(\varepsilon)$, then $\varphi_{x+y_n}(a-\varepsilon) > 0$, which is again a contradiction to the fact that $\mathcal{M}$ is a maximal $L(\alpha)$-nonnegative subspace. ∎

Remark 2.2. The proof of Theorem 2.1 shows that Assumption 3 is needed only if $\sigma(L) \cap (\alpha, \lambda_e) = \emptyset$.

Let $\mathcal{K}$ be a Krein space, and let $\mathcal{K} = \mathcal{H}_+ \oplus \mathcal{H}_-$ be a canonical decomposition of $\mathcal{K}$. We denote the corresponding orthogonal projections onto $\mathcal{H}_\pm$ by $P_\pm$. Each nonnegative subspace $\mathcal{L}$ of $\mathcal{K}$, if it is not maximal nonnegative, is contained in infinitely many maximal nonnegative subspaces. If $\mathcal{M}$ denotes one of these, the dimension of the factor space $\mathcal{M}/\mathcal{L}$ is independent of the choice of the maximal nonnegative subspace $\mathcal{M}$, and it coincides with the dimension of the space $\mathcal{H}_+ \ominus P_+\mathcal{L}$. Moreover, the space $\mathcal{M}_\mathcal{L} := \mathcal{L} \oplus (\mathcal{H}_+ \ominus P_+\mathcal{L})$ is a maximal nonnegative subspace containing $\mathcal{L}$, and $\mathcal{L}$ is the orthogonal complement in $\mathcal{M}_\mathcal{L}$ of the subspace

$$\mathcal{V}_\mathcal{L} := \mathcal{H}_+ \ominus P_+\mathcal{L}. \tag{2.2}$$

The space $\mathcal{M}_\mathcal{L}$ will be called the *standard maximal nonnegative extension* of $\mathcal{L}$.

We also need the following lemma; cf. [M].

LEMMA 2.3. *Suppose that the operator function L, defined on the interval $[\alpha, \beta)$, satisfies Assumptions* 1 *and* 2. *If $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are eigenvalues of L with corresponding eigenvectors $y_1, y_2, \ldots, y_n$ and $\mathcal{L}$ is a subspace of $\mathcal{H}$ such that $p(x) \geq \lambda_n$ for all $x \in \mathcal{L}$, then $(L(\lambda_1)x, x) \geq 0$ for all $x \in \operatorname{span}\{\mathcal{L}, y_1, y_2, \ldots, y_n\}$.*

Proof. If $y \in \mathcal{L}$, then $(L(\lambda_n)(y+y_n), y+y_n) = (L(\lambda_n)y, y) \geq 0$, and hence also $(L(\lambda_{n-1})(y+y_n), y+y_n) \geq 0$. If $n \geq 2$ the same reasoning yields

$$(L(\lambda_{n-1})(y+y_n+y_{n-1}), y+y_n+y_{n-1}) \geq 0,$$

and repeating this we finally get $(L(\lambda_1)x, x) \geq 0$ for $x \in \operatorname{span}\{\mathcal{L}, y_1, y_2, \ldots, y_n\}$, which implies $(L(\alpha)x, x) \geq 0$ for the same x. ∎

THEOREM 2.4. *Suppose that the operator function L, defined on the interval $[\alpha, \beta)$, satisfies Assumptions* 1 *and* 2. *If L has at least n eigenvalues in (α, λ_e) and we denote the n smallest ones by*

$$\lambda_1 = \cdots = \lambda_{n_1} < \lambda_{n_1+1} = \cdots = \lambda_{n_2} < \cdots < \lambda_{n_k+1} = \cdots = \lambda_{n_{k+1}} < \cdots \leq \lambda_n, \tag{2.3}$$

counted according to their multiplicities, then

$$\sup_{\mathcal{M}\in\mathbf{M}_\alpha^+} \sup_{\substack{\mathcal{V}\subset\mathcal{M}\\ \dim\mathcal{V}=j-1}} \inf_{\substack{x\in\mathcal{M}^1\\ x\perp\mathcal{V}}} p(x) \leq \lambda_j, \qquad j = 1, 2, \ldots, n. \tag{2.4}$$

If the total number of eigenvalues of L in (α, λ_e) is finite, say n, and Assumption 3 is satisfied, then the inequality in (2.4) holds also for $j = n+1, n+2, \dots$ if we define $\lambda_{n+1} = \lambda_{n+2} = \cdots = \lambda_e$.

Proof. Let $y_1, y_2, \dots, y_n$ be a system of linearly independent eigenvectors of L corresponding to the eigenvalues in (2.3). We denote the number on the left-hand side of (2.4) by μ_j. Since the sequence of these numbers is nondecreasing, the relation (2.4) will be proved if we show that $\mu_{n_k} \le \lambda_{n_k}$ for $k = 1, 2, \dots$. We prove this by induction with respect to k.

So for $k = 1$ assume that

$$\lambda_1 = \lambda_{n_1} < \inf_{\substack{x \in \mathcal{M}_0^1 \\ x \perp \mathcal{V}_0}} p(x)$$

for some maximal $L(\alpha)$-nonnegative subspace $\mathcal{M}_0$ and some $(n_1 - 1)$-dimensional subspace $\mathcal{V}_0 \subset \mathcal{M}_0$. Then $(L(\lambda_1)x, x) > 0$ for all $x \in \mathcal{V}_0^\perp \cap \mathcal{M}_0$, $x \neq 0$. If y is any nonzero element of the linear span of $y_1, y_2, \dots, y_{n_1}$, it follows that $y \notin \mathcal{V}_0^\perp \cap \mathcal{M}_0$, and hence $\mathcal{V}_0^\perp \cap \mathcal{M}_0$ and the elements $y_1, y_2, \dots, y_{n_1}$ are linearly independent. Further, we obtain for $x \in \mathcal{V}_0^\perp \cap \mathcal{M}_0$ from Lemma 2.3

$$(L(\lambda_1)(x+y), x+y) = (L(\lambda_1)x, x) \ge 0.$$

Because of Assumption 2,

$$(L(\alpha)u, u) \ge 0 \quad \text{for all} \quad u \in \operatorname{span}\{\mathcal{V}_0^\perp \cap \mathcal{M}_0, \mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_{\mathrm{n}_1}\},$$

or, in words, this subspace is $L(\alpha)$-nonnegative. However, this is impossible since the defect of $\mathcal{V}_0^\perp \cap \mathcal{M}_0$ with respect to $\mathcal{M}_0$ and hence with respect to any maximal $L(\alpha)$-nonnegative subspace is $n_1 - 1$.

The proof of the step from k to $k+1$ is similar. Suppose that

$$\mu_{n_k} \le \lambda_{n_k}, \tag{2.5}$$

but $\mu_{n_{k+1}} > \lambda_{n_{k+1}}$. Then there exists a maximal $L(\alpha)$-nonnegative subspace $\mathcal{M}_0$ and an $(n_{k+1} - 1)$-dimensional subspace $\mathcal{V}_0$ of $\mathcal{M}_0$ such that

$$\inf_{\substack{x \in \mathcal{M}_0^1 \\ x \perp \mathcal{V}_0}} p(x) > \lambda_{n_{k+1}}.$$

Since $\big(L(\lambda_{n_{k+1}})y, y\big) = 0$ for all y in the linear span of $y_{n_k+1}, y_{n_k+2}, \dots, y_{n_{k+1}}$, these elements are linearly independent of $\mathcal{V}_0^\perp \cap \mathcal{M}_0$. Consider

$$\mathcal{L} := \operatorname{span}\{\mathcal{V}_0^\perp \cap \mathcal{M}_0, y_{n_k+1}, y_{n_k+2}, \dots, y_{n_{k+1}}\}.$$

By the same argument as above, this is an $L(\alpha)$-nonnegative subspace, and its defect to a maximal $L(\alpha)$-nonnegative subspace is $\dim \mathcal{V}_0 - (n_{k+1} - n_k) = n_k - 1$. With the standard maximal $L(\alpha)$-nonnegative extension $\mathcal{M}_\mathcal{L}$ and the corresponding subspace $\mathcal{V}_\mathcal{L}$ from (2.2) it follows that

$$\begin{aligned}\lambda_{n_{k+1}} = \inf_{x \in \mathcal{L}} p(x) = \inf_{\substack{x \in \mathcal{M}_\mathcal{L}^1 \\ x \perp \mathcal{V}_\mathcal{L}}} p(x) &\le \sup_{\substack{\mathcal{V} \subset \mathcal{M}_\mathcal{L} \\ \dim \mathcal{V} = n_k - 1}} \inf_{\substack{x \in \mathcal{M}_\mathcal{L}^1 \\ x \perp \mathcal{V}}} p(x) \\ &\le \sup_{\mathcal{M} \in \mathbf{M}_\alpha^+} \sup_{\substack{\mathcal{V} \subset \mathcal{M} \\ \dim \mathcal{V} = n_k - 1}} \inf_{\substack{x \in \mathcal{M}^1 \\ x \perp \mathcal{V}}} p(x) = \mu_{n_k} \le \lambda_{n_k},\end{aligned}$$

where the last inequality is a consequence of the induction assumption (2.5). On the other hand, $\lambda_{n_{k+1}} > \lambda_{n_k}$, a contradiction.

In order to prove the last statement of the theorem, assume that for some $l > 0$ we have $\mu_{n+l} > \lambda_e$. Then there exists a maximal $L(\alpha)$-nonnegative subspace $\mathcal{M}_0$ and an $(n+l-1)$-dimensional subspace $\mathcal{V}_0$ of $\mathcal{M}_0$ such that

$$\inf_{\substack{x\in\mathcal{M}_0^1\\ x\perp\mathcal{V}_0}} p(x) > \lambda_e.$$

Choose $\varepsilon > 0$ such that

$$\lambda_n < \lambda_e - \varepsilon < \lambda_e < \lambda_e + \varepsilon < \inf_{\substack{x\in\mathcal{M}_0^1\\ x\perp\mathcal{V}_0}} p(x).$$

As above, Assumption 3 implies for $x \in \mathcal{M}_0$, $x \perp \mathcal{V}_0$ that $(L(\lambda_e)x, x) \geq c\|x\|^2$ with some $c > 0$. Since λ_e belongs to the essential spectrum of L, that is, 0 belongs to the essential spectrum of $L(\lambda_e)$, for each $\eta > 0$ there exists an l-dimensional subspace $\mathcal{L}_l^\eta$ such that $\|L(\lambda_e)\big|_{\mathcal{L}_l^\eta}\| \leq \eta$. As in the proof of Theorem 2.1 it follows that for some subspace $\mathcal{L}_l^\eta$ it holds $(L(\lambda_e - \varepsilon)x, x) > 0$ for all $x \neq 0$, $x \in \operatorname{span}\{\mathcal{V}_0^\perp \cap \mathcal{M}_0, \mathcal{L}_1^\eta\}$. Now Lemma 2.3 implies that the subspace $\mathcal{M}_1 := \operatorname{span}\{\mathcal{V}_0^\perp \cap \mathcal{M}_0, \mathcal{L}_l^\eta, y_1, y_2, \ldots, y_n\}$ is $L(\lambda_1)$-nonnegative, and hence also $L(\alpha)$-nonnegative. On the other hand, the dimension of the factor space $\mathcal{M}_1/(\mathcal{V}_0^\perp \cap \mathcal{M}_0)$ equals $l + n$. This is a contradiction since $\dim \mathcal{M}_0/(\mathcal{V}_0^\perp \cap \mathcal{M}_0) = n + l - 1$ and the dimensions of all the complementary spaces of $\mathcal{V}_0^\perp \cap \mathcal{M}_0$ to maximal $L(\alpha)$-nonnegative spaces coincide. □

3. Triple variational principles.

3.1. Self-adjoint operators in Hilbert space with a gap in the essential spectrum. In this subsection a self-adjoint operator A on some Hilbert space $\mathcal{H}$ will be considered for which there exists a semiclosed interval such that $\sigma(A)$ is discrete in this interval. The eigenvalues of A in this interval are characterized by a triple variational principle. By $\mathbf{M}_\alpha^+$ we denote the set of maximal $(A - \alpha I)$-nonnegative subspaces; that is, a subspace $\mathcal{M}$ of $\mathcal{H}$ belongs to $\mathbf{M}_\alpha^+$ if $((A - \alpha)x, x) \geq 0$ for all $x \in \mathcal{M}$ and $\mathcal{M}$ is maximal with respect to this property. The results of section 2 will be applied to the linear pencil $L(\lambda) = A - \lambda I$. Evidently, if $\alpha \in \rho(A)$, then L satisfies Assumptions 1–3.

THEOREM 3.1. *Let A be a self-adjoint operator A such that for some $\alpha \in \rho(A) \cap R$ and $\beta > \alpha$ the spectrum $\sigma(A)$ is discrete in the interval $[\alpha, \beta)$. If A has at least n eigenvalues in $[\alpha, \beta)$ and we denote the smallest n ones by*

$$\lambda_1 = \lambda_2 = \cdots = \lambda_{n_1} < \lambda_{n_1+1} = \cdots = \lambda_{n_2} < \cdots < \lambda_{n_k+1} = \cdots = \lambda_{n_{k+1}} < \cdots \leq \lambda_n,$$

counted according to their multiplicities, then

$$\lambda_j = \max_{\mathcal{M}\in\mathbf{M}_\alpha^+} \sup_{\substack{\mathcal{V}\subset\mathcal{M}\\ \dim\mathcal{V}=j-1}} \inf_{\substack{x\in\mathcal{M}^1\\ x\perp\mathcal{V}}} p(x), \qquad j = 1, 2, \ldots, n. \tag{3.1}$$

If A has finitely many, say n eigenvalues in $[\alpha, \beta)$ and $\sigma_{ess}(A) \cap [\alpha, \beta] \neq \emptyset$, then with $\lambda_e = \min \sigma_{ess}(A) \cap [\alpha, \beta]$ and $\lambda_{n+1} = \lambda_{n+2} = \cdots = \lambda_e$ the relation (3.1) holds also for $j = n + 1, n + 2, \ldots$.

Proof. We apply Theorem 2.4 to the linear pencil

$$L(\lambda) := A - \lambda I, \qquad \lambda \in R, \tag{3.2}$$

and obtain the relation (3.1) with $\geq$ instead of the sign $=$. It remains to find a subspace $\mathcal{M}_0 \in \mathbf{M}_\alpha^+$ such that

$$\lambda_j = \max_{\substack{\mathcal{V}\subset\mathcal{M}_0 \\ \dim\mathcal{V}=j-1}} \inf_{\substack{x\in\mathcal{M}_0^1 \\ x\perp\mathcal{V}}} p(x). \tag{3.3}$$

If we choose for $\mathcal{M}_0$ the spectral subspace of A corresponding to the interval $[\alpha,\infty)$ the relation (3.3) is a consequence of a classical double variational principle of the spectrum of a self-adjoint operator, applied to the operator $A|_{\mathcal{M}_0}$. □

Remark 3.2. In Theorem 3.1 the set $\mathbf{M}_\alpha^+$ can be replaced by the set of all maximal $(A-\alpha I)$-positive subspaces.

3.2. Nonnegative operators on Krein spaces. Let $(\mathcal{K}, [\,\cdot\,,\,\cdot\,])$ be a Krein space. We fix a fundamental symmetry J on $\mathcal{K}$ and introduce the Hilbert space $(\mathcal{H}, (\,\cdot\,,\,\cdot\,))$, which consists of the same elements as $\mathcal{K}$ and with inner product

$$(x,y) := [Jx,y], \qquad x,y\in\mathcal{K}.$$

Let A be a bounded positive operator on $\mathcal{K}$; here positive means that $[Ax,x]>0$ for all $x\in\mathcal{K}$, $x\neq 0$. Then the spectrum $\sigma(A)$ is real, with positive eigenvalues having positive-type and negative eigenvalues having negative-type eigenvectors; see, e.g. [L], [AI]. The spectrum of A on $\mathcal{K}$ and also the eigenvalues, eigenvectors, etc. coincide with the spectrum, the eigenvalues, etc. of the self-adjoint linear pencil $M(\mu) := JA-\mu J$ on the Hilbert space $\mathcal{H}$.

We introduce the pencil

$$L(\lambda) := J-\lambda JA.$$

The spectrum of the pencil L is in general unbounded, the relations

$$\lambda\in\sigma(L) \Longleftrightarrow \mu=\frac{1}{\lambda}\in\sigma(M), \qquad \lambda\in\sigma_p(L) \Longleftrightarrow \mu=\frac{1}{\lambda}\in\sigma_p(M)$$

hold, and the eigenvectors of corresponding eigenvalues of L and M coincide. The pencil L satisfies Assumptions 1–2 with respect to $\alpha=0$ and $\beta=\infty$. Denote by $\mathbf{M}_0^+$ the set of all maximal nonnegative subspaces of $\mathcal{K}$, by $\mathbf{M}_0^{++}$ the set of all maximal positive subspaces of $\mathcal{K}$.

THEOREM 3.3. *Let A be a positive operator on the Krein space $\mathcal{K}$. Suppose that A has at least n eigenvalues which are greater than* $\max\{\sigma_{ess}(A),0\}$ *denoted by*

$$\mu_1=\mu_2=\cdots=\mu_{n_1}>\mu_{n_1+1}=\cdots=\mu_{n_2}>\cdots>\mu_{n_k+1}=\cdots=\mu_{n_{k+1}}>\cdots\geq\mu_n,$$

counted according to their multiplicities. Then

$$\mu_j = \inf_{\mathcal{M}\in\mathbf{M}_0^{++}} \inf_{\substack{\mathcal{V}\subset\mathcal{M} \\ \dim\mathcal{V}=j-1}} \sup_{\substack{x\in\mathcal{M}^1 \\ x\perp\mathcal{V}}} \frac{[Ax,x]}{[x,x]}, \qquad j=1,2,\dots,n. \tag{3.4}$$

Proof. If we apply Theorem 2.4 to the pencil L and the interval $[0,\infty)$ we obtain with $\lambda_j=\mu_j^{-1}$, $j=1,2,\dots,n$,

$$\lambda_j \geq \sup_{\mathcal{M}\in\mathbf{M}_0^+} \sup_{\substack{\mathcal{V}\subset\mathcal{M} \\ \dim\mathcal{V}=j-1}} \inf_{\substack{x\in\mathcal{M}^1 \\ x\perp\mathcal{V}}} \frac{[x,x]}{[Ax,x]}, \qquad j=1,2,\dots,n.$$

Rewriting this relation for the μ_j and positive subspaces $\mathcal{M}$ gives

$$\mu_j \le \inf_{\mathcal{M}\in\mathbf{M}_0^{++}} \inf_{\substack{\mathcal{V}\subset\mathcal{M}\\ \dim\mathcal{V}=j-1}} \sup_{\substack{x\in\mathcal{M}^1\\ x\perp\mathcal{V}}} \frac{[Ax,x]}{[x,x]}, \qquad j=1,2,\dots,n.$$

It remains to find a subspace $\mathcal{M}_0 \in \mathbf{M}_0^{++}$ such that

$$\mu_j = \inf_{\substack{\mathcal{V}\subset\mathcal{M}_0\\ \dim\mathcal{V}=j-1}} \sup_{\substack{x\in\mathcal{M}_0^1\\ x\perp\mathcal{V}}} \frac{[Ax,x]}{[x,x]}, \qquad j=1,2,\dots,n.$$

We choose $\mathcal{M}_0$ to be a maximal nonnegative subspace of $\mathcal{K}$ which is invariant under A and which exists according to [L, Theorem 7.1]. Then $\mathcal{M}_0$ is even a positive subspace: If it contains a neutral element x_0 it follows that $[Ax_0, x_0] = 0$, and hence $x_0 = 0$ since A is positive. Now consider the Hilbert space completion $\mathcal{H}_0$ of $\mathcal{M}_0$ with respect to the inner product $[\,\cdot\,,\,\cdot\,]$. The restriction $A|_{\mathcal{M}_0}$ extends by continuity to a bounded self-adjoint operator A_0 in $\mathcal{H}_0$ which has the same discrete spectrum as the restriction $A|_{\mathcal{M}_0}$. If we apply the classical variational principle to A_0 the claim follows. □

3.3. A class of quadratic operator pencils. In this subsection we consider a self-adjoint quadratic operator pencil

$$L(\lambda) = -\lambda^2 I + \lambda B + C$$

with bounded operators B and C in some Hilbert space $\mathcal{H}$, B being nonpositive. As we have mentioned already, pencils of this form with unbounded operators arise in problems of mechanics; see [Pi1], [Pi2]. Here, however, we shall restrict ourselves to the case of bounded operators.

For convenience it is also assumed that C is boundedly invertible: $0 \in \rho(C)$, and we write C as the difference of its two positive components C_+, C_- : $C = C_+ - C_-$. We shall characterize the smallest discrete positive eigenvalues of L by a variational principle from the left.

The pencil L satisfies Assumptions 1–3 with respect to $\alpha = 0$ and $\beta = \infty$. Denote by $\mathbf{M}_0^+$ the set of all C-nonnegative subspaces of $\mathcal{H}$. For $x \neq 0$ the solutions of the equation

$$\lambda^2\|x\|^2 - \lambda(Bx,x) - (Cx,x) = 0$$

are

$$\lambda = p_\pm(x) = \frac{1}{2\|x\|^2}\left((Bx,x) \pm \sqrt{(Bx,x)^2 + 4(Cx,x)\|x\|^2}\right).$$

Hence this equation has a solution in the right half plane if and only if $(Cx,x) > 0$, and then this solution $p_+(x)$ is unique and real; we denote it for short by $p(x)$. It follows that under the above assumptions the spectrum of L in the right half plane is real.

THEOREM 3.4. *Given the quadratic operator pencil*

$$L(\lambda) = -\lambda^2 I + \lambda B + C$$

with a bounded nonpositive operator B and a bounded self-adjoint operator C such that $0 \in \rho(C)$. Then the spectrum of L in the right half plane is real, and hence

positive, and it is nonempty if and only if $C_+ \neq 0$. *If* L *has at least* n *eigenvalues in the interval* $[0, \lambda_e)$, *where* $\lambda_e := \min \sigma_{ess}(L) \cap R^+$ *and we denote the* n *smallest ones as in* (2.3):

$$\lambda_1 = \cdots = \lambda_{n_1} < \lambda_{n_1+1} = \cdots = \lambda_{n_2} < \cdots < \lambda_{n_k+1} = \cdots = \lambda_{n_{k+1}} < \cdots \leq \lambda_n,$$

counted according to their multiplicities, then

$$\sup_{\mathcal{M} \in \mathbf{M}_0^+} \sup_{\substack{\mathcal{V} \subset \mathcal{M} \\ \dim \mathcal{V} = j-1}} \inf_{\substack{x \in \mathcal{M}^1 \\ x \perp \mathcal{V}}} p(x) = \lambda_j, \qquad j = 1, 2, \ldots, n. \tag{3.5}$$

If the total number of eigenvalues of L *in* $[0, \lambda_e)$ *is finite, say* n, *and* $\operatorname{ran} C_+$ *is infinite-dimensional, then the equality sign in* (3.5) *holds also for* $j = n+1, n+2, \ldots$ *if we define* $\lambda_{n+1} = \lambda_{n+2} = \cdots = \lambda_e$.

Proof. Theorem 2.4 yields immediately the inequalities

$$\sup_{\mathcal{M} \in \mathbf{M}_\alpha^+} \sup_{\substack{\mathcal{V} \subset \mathcal{M} \\ \dim \mathcal{V} = j-1}} \inf_{\substack{x \in \mathcal{M}^1 \\ x \perp \mathcal{V}}} p(x) \leq \lambda_j, \qquad j = 1, 2, \ldots, n. \tag{3.6}$$

We shall find a maximal C-positive subspace $\mathcal{M}_0$ such that

$$\sup_{\substack{\mathcal{V} \subset \mathcal{M}_0 \\ \dim \mathcal{V} = j-1}} \inf_{\substack{x \in \mathcal{M}_0^1 \\ x \perp \mathcal{V}}} p(x) = \lambda_j, \qquad j = 1, 2, \ldots, n.$$

In order to find such a subspace we consider the following linearization of the pencil L. In the space $\widetilde{\mathcal{H}} := \mathcal{H} \oplus \mathcal{H}$ define the operators

$$\widetilde{A} := \begin{pmatrix} B & C \\ I & 0 \end{pmatrix}, \qquad \widetilde{G} := \begin{pmatrix} -I & 0 \\ 0 & C \end{pmatrix}. \tag{3.7}$$

The operator $\widetilde{A}$ is a standard linearization of the pencil L, and it is well known that the spectra of L and of $\widetilde{A}$ coincide. In particular, the spectrum of $\widetilde{A}$ in the right half plane forms a spectral set in the Riesz–Dunford sense. Denote the corresponding spectral subspace of $\widetilde{A}$ by $\widetilde{\mathcal{M}}_0$. The operator $\widetilde{A}$ is $\widetilde{G}$-accretive:

$$\Re\,(\widetilde{G}\widetilde{A}) = \Re \begin{pmatrix} -B & -C \\ C & 0 \end{pmatrix} = \begin{pmatrix} -B & 0 \\ 0 & 0 \end{pmatrix} \geq 0.$$

Therefore this spectral subspace $\widetilde{\mathcal{M}}_0$ is $\widetilde{G}$-nonnegative; see [AI]. It follows that it admits a representation of the form

$$\widetilde{\mathcal{M}}_0 = \left\{ \begin{pmatrix} K P_{\mathcal{M}_0} x \\ P_{\mathcal{M}_0} x \end{pmatrix} : \, x \in \mathcal{H} \right\},$$

where $P_{\mathcal{M}_0}$ is an orthogonal projection in $\mathcal{H} = \operatorname{ran} C_- \oplus \operatorname{ran} C_+$ onto a subspace $\mathcal{M}_0$ of the form

$$\mathcal{M}_0 = \left\{ \begin{pmatrix} K_1 x \\ x \end{pmatrix} : \, x \in \operatorname{ran} C_+ \right\},$$

with a bounded linear operator K_1 from $\operatorname{ran} C_+$ into $\operatorname{ran} C_-$. It is now easy to see that the spectrum of the operator $\widetilde{A}|_{\widetilde{\mathcal{M}}_0}$, which is the spectrum of $\widetilde{A}$ in the right half plane, coincides with the spectrum of the pencil

$$L_{\mathcal{M}_0}(\lambda) := \lambda^2 P_{\mathcal{M}_0} - \lambda P_{\mathcal{M}_0} B P_{\mathcal{M}_0} - P_{\mathcal{M}_0} C P_{\mathcal{M}_0},$$

(which is considered just in the subspace $\mathcal{M}_0$ of $\mathcal{H}$) in the right half plane, and, in particular, the eigenvalues in the right half plane and the corresponding eigenvectors of the pencils L and $L_{\mathcal{M}_0}$ coincide. On the other hand, we have from (3.7) for $\widetilde{x} = \binom{KP_{\mathcal{M}_0}x}{P_{\mathcal{M}_0}x} \in \widetilde{\mathcal{M}}_0$

$$(\widetilde{G}\widetilde{x}, \widetilde{x}) = (CP_{\mathcal{M}_0}x, P_{\mathcal{M}_0}x) - \|KP_{\mathcal{M}_0}x\|^2 \geq 0,$$

and it follows that $P_{\mathcal{M}_0}CP_{\mathcal{M}_0}$ is strictly positive on $\mathcal{M}_0$, or, in other words, the subspace $\mathcal{M}_0$ is $P_{\mathcal{M}_0}CP_{\mathcal{M}_0}$-positive. Therefore the positive eigenvalues of the pencil $L_{\mathcal{M}_0}$ can be characterized by double variational principles, which implies that

$$\sup_{\substack{\mathcal{V}\subset\mathcal{M}_0\\ \dim\mathcal{V}=j-1}} \inf_{\substack{x\in\mathcal{M}_0{}^1\\ x\perp\mathcal{V}}} p(x) = \lambda_j, \qquad j = 1, 2, \ldots, n;$$

that is, for $\mathcal{M} = \mathcal{M}_0$ the equality sign in (3.6) is attained. □

3.4. Block operator matrices. In this subsection we consider a self-adjoint operator $\widetilde{A}$ on the orthogonal sum $\widetilde{\mathcal{H}} = \mathcal{H}_1 \oplus \mathcal{H}_2$ of two Hilbert spaces $\mathcal{H}_1$, $\mathcal{H}_2$ given by the block operator matrix

$$\widetilde{A} = \begin{pmatrix} A & B \\ B^* & D \end{pmatrix}. \tag{3.8}$$

Evidently, A and D are self-adjoint operators on $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. The spectrum of $\widetilde{A}$ outside of $\sigma(D)$ coincides with the spectrum of the first Schur complement

$$L(\lambda) := A - \lambda I - B(D - \lambda I)^{-1}B^*$$

of $\widetilde{A}$. If $\lambda \notin \sigma(D)$, then $L'(\lambda) = -I - B(D - \lambda I)^{-2}B^* \leq -I$; therefore, if $\alpha > \max\sigma(D)$, $\alpha \in \rho(\widetilde{A})$, and $\beta > \alpha$, then for L and the interval $[\alpha, \beta)$ Assumptions 1–3 are satisfied. In particular, for $x \in \mathcal{H}_1$, $x \neq 0$, the equation $(L(\lambda)x, x) = 0$ has at most one zero in the interval $[\alpha, \infty)$ and has exactly one zero in this interval if $(L(\alpha)x, x) \geq 0$. Denote this zero by $p(x)$. Further, by $\mathbf{M}_\alpha^+$ we denote the set of all maximal $L(\alpha)$-nonnegative subspaces of $\mathcal{H}_1$.

THEOREM 3.5. *Let the self-adjoint block operator matrix $\widetilde{A}$ be given as in* (3.8). *Consider $\alpha \in \rho(\widetilde{A})$ such that $\alpha > \max\sigma(D)$, and denote $\lambda_e := \min\sigma_{ess}(\widetilde{A}) \cap [\alpha, \infty)$. If $\widetilde{A}$ has at least n eigenvalues in (α, λ_e) and we denote the n smallest ones as in* (2.3):

$$\lambda_1 = \cdots = \lambda_{n_1} < \lambda_{n_1+1} = \cdots = \lambda_{n_2} < \cdots < \lambda_{n_k+1} = \cdots = \lambda_{n_{k+1}} < \cdots \leq \lambda_n,$$

counted according to their multiplicities, then

$$\lambda_j = \sup_{\mathcal{M}\in\mathbf{M}_\alpha^+} \sup_{\substack{\mathcal{V}\subset\mathcal{M}\\ \dim\mathcal{V}=j-1}} \inf_{\substack{x\in\mathcal{M}^1\\ x\perp\mathcal{V}}} p(x), \qquad j = 1, 2, \ldots, n. \tag{3.9}$$

If the total number of eigenvalues of L in (α, λ_e) is finite, say n, then the equality in (3.9) *holds also for $j = n+1, n+2, \ldots$ if we define $\lambda_{n+1} = \lambda_{n+2} = \cdots = \lambda_e$.*

Proof. Theorem 2.4 yields immediately the relation (3.9) with the sign $\geq$ instead of the sign $=$. It remains to find a maximal $L(\alpha)$-nonnegative subspace of $\mathcal{H}_1$ such that the equality is attained. To this end we consider the spectral invariant subspace $\widetilde{\mathcal{M}}$

of $\widetilde{A}$ corresponding to the interval $\Delta := [\alpha, \infty)$. According to [LMMT, Theorem 1.2] this invariant subspace is of the form

$$\widetilde{\mathcal{M}} = \left\{ \begin{pmatrix} x \\ K_\Delta x \end{pmatrix} : x \in \mathcal{H}_1^\Delta \right\},$$

where $\mathcal{H}_1^\Delta$ is a subspace of $\mathcal{H}_1$ and K_Δ is a bounded operator from $\mathcal{H}_1^\Delta$ into $\mathcal{H}_2$. Let P_Δ be the orthogonal projection in $\mathcal{H}_1$ onto $\mathcal{H}_1^\Delta$. We introduce the space $\widetilde{\mathcal{H}}^\Delta := \mathcal{H}_1^\Delta \oplus \mathcal{H}_2$ and the compression $\widetilde{A}_\Delta$ of $\widetilde{A}$ to $\widetilde{\mathcal{H}}^\Delta$:

$$\widetilde{A}_\Delta := \begin{pmatrix} P_\Delta A P_\Delta & P_\Delta B \\ B^* P_\Delta & D \end{pmatrix}.$$

Evidently, with the first Schur complement $L_\Delta(\lambda)$ of $\widetilde{A}_\Delta$ it holds that

$$(L(\lambda)x, x) = (L_\Delta(\lambda)x, x), \qquad x \in \mathcal{H}_1^\Delta.$$

Then, according to [LMMT, Theorem 2.6], $\mathcal{M}_0 := \mathcal{H}_1^\Delta$ is a maximal $L(\alpha)$-positive subspace. The discrete eigenvalues of $\widetilde{A}$ and of $\widetilde{A}_\Delta$ in $[\alpha, \infty)$ coincide, and it is well known (see [BEL]) that the latter can be characterized by the formula

$$\lambda_j = \sup_{\substack{\mathcal{V} \subset \mathcal{M}_0 \\ \dim \mathcal{V} = j-1}} \inf_{\substack{x \in \mathcal{M}_0^1 \\ x \perp \mathcal{V}}} p(x), \qquad j = 1, 2, \ldots, n. \qquad \square$$

REFERENCES

[ALM] V. Adamyan, H. Langer, and R. Mennicken, *Eigenvalues of a Sturm-Liouville problem depending rationally on the eigenvalue parameter*, Math. Res. 79, Akademie Verlag, Berlin, 1994, pp. 589–594.

[AI] T. Ya. Azizov and I. S. Iokhvidov, *Linear Operators in Spaces with an Indefinite Metric*, John Wiley and Sons, Chichester, UK, 1989.

[B] J. Bognár, *Indefinite Inner Product Spaces*, Springer-Verlag, Berlin, 1974.

[BEL] P. Binding, D. Eschwé, and H. Langer, *Variational principles for real eigenvalues of self-adjoint operator pencils*, Integral Equations Operator Theory, 38 (2000), pp. 190–206.

[GK] I. C. Gohberg and M. G. Kreĭn, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monogr. 18, AMS, Providence, RI, 1969.

[GS] M. Griesemer and H. Siedentop, *A minimax principle for the eigenvalues in spectral gaps*, J. London Math. Soc. (2), 60 (1999), pp. 490–500.

[L] H. Langer, *Spectral functions of definitizable operators in Krein spaces*, in Proceedings of the Graduate School "Functional Analysis", Dubrovnik, 1981, Lecture Notes in Math. 948, Springer-Verlag, Berlin, 1982, pp. 1–46.

[LMMT] H. Langer, A. Markus, V. Matsaev, and C. Tretter, *Self-adjoint block operator matrices with non-separated diagonal entries*, J. Funct. Anal., to appear.

[LM] H. Langer and M. Möller, *The essential spectrum of a non-elliptic boundary value problem*, Math. Nachr., 178 (1996), pp. 233–248.

[M] A. S. Markus, *Introduction to the Spectral Theory of Polynomial Operator Pencils*, Transl. Math. Monogr. 71, AMS, Providence, RI, 1988.

[P] R. S. Phillips, *A minimax characterization for the eigenvalues of a positive symmetric operator in a space with an indefinite metric*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 17 (1970), pp. 51–59.

[Pi1] V. N. Pivovarchik, *Eigenvalues of a quadratic operator pencil*, Funct. Anal. Appl., 23 (1989), pp. 70–72.

[Pi2] V. N. Pivovarchik, *Oscillations of a semi-infinite rod with internal and external friction*, J. Appl. Math. Mech., 52 (1988), pp. 647–653.

[T] B. Textorius, *Minimaxprinzipe zur Bestimmung der Eigenwerte J-nichtnegativer Operatoren*, Math. Scand., 35 (1974), pp. 105–114.

SIAM J. MATH. ANAL.
Vol. 34, No. 1, pp. 239–256

CRITICAL MAGNETIC FIELD AND ASYMPTOTIC BEHAVIOR OF SUPERCONDUCTING THIN FILMS*

SHIJIN DING† AND QIANG DU‡

Abstract. In this paper, we discuss the vortex structure of the superconducting thin films placed in a magnetic field. The discussion is based on a system of simplified Ginzburg–Landau equations. We obtain the estimate for the lower critical magnetic field H_{c_1}, in the sense that it is the first critical value of h_{ex}, the applied field, for which the minimal energy among vortexless configurations is equal to the minimal energy among single-vortex configurations; moreover, it corresponds to the first phase transition in which vortices appear in the superconductor. We also discuss the location of these vortices and the asymptotic behavior of the local minimizers.

Key words. superconductivity, thin films, vortices, pinning, critical magnetic field

AMS subject classifications. 35J55, 35Q40

PII. S0036141000378619

1. Introduction.

Consider a three-dimensional superconducting thin film that occupies the domain $\Omega_\delta = \Omega \times (-\delta a, \delta a)$, where Ω is a bounded smooth planar domain and $a \in C^\infty(\overline{\Omega})$ is a function measuring the variation in the film thickness. Assume that $a(x) \geq a_0 > 0$ for all $x \in \overline{\Omega}$; by taking integral averages along the vertical direction and setting δ going to zero, it was shown in [10] that the three-dimensional Ginzburg–Landau model of superconductivity [16, 26] defined on Ω_δ may be reduced to a two-dimensional one given by the minimization in $H^1(\Omega)$ of the functional

$$J_a(u) = \frac{1}{2}\int_\Omega a(x)\left[|\nabla_{\mathbf{A}_0} u|^2 + \frac{1}{2\varepsilon^2}(1-|u|^2)^2\right], \tag{1.1}$$

where $\mathbf{A}_0(x)$, the in-plane component of the magnetic potential, is determined by

$$\begin{cases} \operatorname{div}(a(x)\mathbf{A}_0) = 0, \quad \mathbf{curl}\mathbf{A}_0 = h_{ex} & \text{in } \Omega, \\ \mathbf{A}_0 \cdot \mathbf{n} = \mathbf{0} & \text{on } \partial\mathbf{\Omega}. \end{cases} \tag{1.2}$$

Here, h_{ex} is the external magnetic field which is applied vertically to the (x_1, x_2)-plane, $\mathbf{n}$ denotes the outward normal to $\partial\Omega$, u is the complex superconducting order parameter with $|u|^2$ representing the density of superconducting electrons ($|u| = 1$ corresponds to the superconducting state, $|u| = 0$ corresponds to the normal state), $\nabla_{\mathbf{A}_0} u = \nabla u - i\mathbf{A}_0 u$, and ε is proportional to the coherence length.

Let u be a critical point of the functional $J_a(u)$ in $H^1(\Omega)$ which satisfies the

*Received by the editors September 20, 2000; accepted for publication (in revised form) March 9, 2002; published electronically September 24, 2002. This work has been partially supported by the State Key Basic Research Project G199903280, the Natural Science Foundation of China (19971030), the Natural Science Foundation of Guangdong Province (000671), and by grant DMS-0196522 from US NSF.

http://www.siam.org/journals/sima/34-1/37861.html

†Department of Mathematics, South China Normal University, Guangzhou, Guangdong 510631, People's Republic of China (dingsj@scnu.edu.cn).

‡Department of Mathematics, Penn State University, University Park, PA 16802, and Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong (qdu@math.psu.edu).

Euler–Lagrange (or simplified Ginzburg–Landau) equation

$$\begin{cases} -(\nabla - i\mathbf{A}_0)\cdot a(x)(\nabla u - i\mathbf{A}_0 u) = \dfrac{a(x)}{\varepsilon^2}u(1-|u|^2) \text{ in } \Omega, \\ \partial_{\mathbf{n}} u = 0 \qquad\qquad\qquad \text{on } \partial\Omega. \end{cases} \tag{1.3}$$

The points where the zeros of u appear, with their topological degrees, are called the vortices of the map u. Understanding the vortex structures in the solutions and describing the vortices as h_{ex} varies is of great physical relevance and mathematical interests. Discussions on the vortex state in the thin film geometry have been given in [1, 16, 17, 19, 20, 26]; in particular, the variation in the film thickness is thought to provide an effective vortex pinning mechanism [10]. For works related to the mathematical analysis of the various pinning mechanisms, we refer to [2, 3, 4, 6, 10, 11, 12, 15].

In [7, 8], rigorous mathematical analysis of vortex solutions has been done for a similar problem with $a(x) = 1$, $\mathbf{A}_0 = 0$ and Dirichlet boundary condition $u = g\colon \Omega \to S^1$ of degree d. It was proved that, asymptotically, minimizers have d isolated vortices of degree one and their locations are determined by minimizing a renormalized energy. This result was extended to the case $a(x) \not\equiv 1$, $\mathbf{A}_0 = 0$ with the same Dirichlet boundary conditions in [6] and [15] independently, and the vortices of the minimizers were shown to be located at the minimum of $a(x)$. Some results similar to those in [7] were obtained in [9] for the original Ginzburg–Landau functional $J(u, \mathbf{A})$,

$$J(u,\mathbf{A}) = \frac{1}{2}\int_\Omega \left[|\nabla_{\mathbf{A}} u|^2 + |\mathbf{curlA} - h_{ex}|^2 + \frac{1}{2\varepsilon^2}(1-|u|^2)^2\right],$$

with $h_{ex} = 0$ and the gauge invariant Dirichlet conditions (a name given in [22]). This work was later extended in [14] to the case where a weight (thickness) appears in the functional $J(u, \mathbf{A})$; the corresponding renormalized energy was presented in [13]. Similar analysis based on the functional (1.1) was also presented in [18]. All the available results substantiate the pinning effect of the thickness variation; that is, the vortices turn to stay where the film is thin.

Recently, the minimizers of $J(u, \mathbf{A})$ with nonzero applied fields with natural boundary conditions were studied in [5, 18, 23, 24, 25, 21, 22]. In this case, there is no a priori bound on the number of the vortices for the minimizers in $H^1 \times H^1$. To overcome this difficulty, i.e., to have an a priori control on the numbers of the vortices, in [23, 24], the local minimizers of the functional

$$J(u,\mathbf{A}) = \frac{1}{2}\int_\Omega \left[|\nabla_{\mathbf{A}} u|^2 + |\mathbf{curlA} - h_{ex}|^2 + \frac{1}{2\varepsilon^2}(1-|u|^2)^2\right]$$

in the set $\overline{D_M}$ were studied, where

$$D_M = \left\{(u,\mathbf{A}) \in H^1(\Omega)\times H^1(\Omega) :\ F(u) < M|\ln\varepsilon|\ \right\}$$

and $F(u) = J_1(u) = J(u, 0)$ with $\mathbf{A}_0 = 0$. The minimizers were shown not to be on the boundary of D_M, hence the Ginzburg–Landau equations (the Euler–Lagrange equations for the functional J) are satisfied. Such analysis also provided estimates on the lower critical magnetic field H_{c_1}, the locations of the vortices, and the asymptotic behaviors of the minimizers. The lower critical field H_{c_1} may be defined as the value of h_{ex} for which the minimal energy among vortexless configurations is equal to the

minimal energy among single-vortex configurations. For $h_{ex} \le H_{c_1}$, it was shown in [21] that the global minimizer (in $H^1 \times H^1$) of Ginzburg–Landau functional $J(u, \mathbf{A})$ is the vortexless solution found in [23]. For the case $H_{c_1} \ll h_{ex} \ll H_{c_2}$, in [22], it was shown that as $\varepsilon \to 0$ the energy minimizers have vortices whose density tends to be uniform and proportional to h_{ex}. For other discussions, we refer the reader to [2] and [25] and references therein.

In this paper, we study the minimizers of the functional (1.1) in the set

$$D_M^a = \left\{ u \in H^1(\Omega) : \ F_a(u) < M|\ln\varepsilon| \ \right\}, \tag{1.4}$$

where $F_a(u) = J_a(u)$ with $\mathbf{A}_0 = 0$. The main techniques of this paper come from [23, 24]. We also present the estimate on the lower critical magnetic field H_{c_1} and discuss the impact of the thickness function $a(x)$ and the given applied field $\mathbf{curl}\mathbf{A}_0$ on the vortices: their number and their locations. These new results have not been stated even in the physics literature. Our results also provide rigorous theoretical justification of the pinning mechanism due to the thickness variation based on the simplified Ginzburg–Landau model.

Let us introduce a few notation. By (1.2), there is a function $\xi \in H^2(\Omega)$ such that

$$a(x)\mathbf{A}_0(x) = \nabla^\perp \xi = (-\xi_{x_2}, \xi_{x_1}) \ \text{ in } \Omega.$$

Using the scaling $\xi = \xi_0 h_{ex}$, we have from (1.2) that

$$\begin{cases} -\mathrm{div}(\frac{1}{a(x)}\nabla\xi_0) = -1 & \text{in } \Omega, \\ \xi_0 = 0 & \text{on } \partial\Omega. \end{cases} \tag{1.5}$$

By the maximum principle, we may easily see that $-C \le \xi_0 < 0$ for some constant $C > 0$ and ξ_0 is a smooth function that depends only on Ω and $a = a(x)$. Let

$$\Lambda = \left\{ x \in \Omega, |\xi_0(x)/a(x)| = \max_{y\in\Omega} |\xi_0(y)/a(y)| \right\}. \tag{1.6}$$

To state our main results, the following assumption is made.

Assumption 1.1. *Assume that the constant M in* (1.4) *is chosen so that there is a positive integer $n \in \mathbb{N}$ such that*

$$\left[\frac{M}{\pi \max\limits_{\Lambda} a(x)}, \quad \frac{M}{\pi \min\limits_{\Lambda} a(x)} \right] \subset (n, n+1). \tag{1.7}$$

The above assumption on the existence of $n \in \mathbb{N}$ with the desired property (1.7) is needed in proving (see section 6) that the minimizer of $J_a(u)$ in $\overline{D_M^a}$ is in D_M^a (not on ∂D_M^a) and thus the minimizer is a solution of (1.3). Under the above assumption, we have the following theorem.

Theorem 1.1. *There exists $k_a = \frac{1}{2\max_\Omega |\xi_0(x)/a(x)|}$, $k_2^\varepsilon = O(1)$, $k_3^\varepsilon = o(1)$, and $\varepsilon_0 = \varepsilon_0(M) > 0$ such that*

$$H_{c_1} = k_a|\ln\varepsilon| + k_2^\varepsilon, \tag{1.8}$$

and, for $\varepsilon < \varepsilon_0$, the following holds:

(i) *If* $h_{ex} \le H_{c_1}$, *there exists a solution* u_ε *of* (1.3) *which minimizes* $J_a(u)$ *in* D_M^a, *and it satisfies* $1/2 \le |u_\varepsilon| \le 1$.

(ii) *If* $H_{c_1} + k_3^\varepsilon \le h_{ex} \le H_{c_1} + O(1)$, *there exists a solution* u_ε *of* (1.3) *that minimizes* $J_a(u)$ *in* D_M^a. *The solution has a bounded positive number of vortices* b_i^ε *of degree one such that*

$$\operatorname{dist}(b_i^\varepsilon, \Lambda) \to 0, \quad \text{as } \varepsilon \to 0, \tag{1.9}$$

and there exists a constant $\alpha > 0$ *such that* $\operatorname{dist}(b_i^\varepsilon, b_j^\varepsilon) \ge \alpha$ *for* $i \ne j$.

REMARK 1.1. *The main differences between our results and those in* [23, 24] *are as follows: first,* A_0 *is determined a priori, and it satisfies* (1.2) *and* $\mathbf{curl} A_0(x) = O(|\ln \varepsilon|)$ *so that no London type equation is used; second, with a variable weight* $a = a(x)$ *in the functional, methods developed in* [6] *(see also* [13]*) and in* [23] *are needed to derive the energy lower bound.*

REMARK 1.2. *It follows from the proof of Theorem* 1.1 *that the number of the vortices, under our assumption, is bounded by*

$$N = \min\left\{ \frac{M}{\pi \max_\Lambda a(x)}, \quad \frac{\min_\Lambda a(x)}{\max_\Lambda a(x) - \min_\Lambda a(x)} \right\}.$$

REMARK 1.3. *From* (1.6) *and* (1.9)*, one may conclude that the distribution of the vortex locations are influenced both by the pinning effect due to thickness variation and the effect of the applied magnetic field. A similar phenomenon has also been explored in* [2] *with normal inclusion serving as pinning sites.*

Let us discuss briefly Assumption 1.1 and the results of Theorem 1.1. The parameter M in (1.4) is chosen such that

$$\Big[M/(\pi \max_\Lambda a(x)), M/(\pi \min_\Lambda a(x)) \Big] \subset (n, n+1)$$

for some positive integer n. For Λ defined by (1.6), it is easy to see that the above assumption can be equivalently replaced by

$$\max_\Lambda a(x) < 2 \min_\Lambda a(x). \tag{1.10}$$

Note that if Λ consists of only one point, or if $a(x)$ satisfies

$$\max_{\overline{\Omega}} a(x) < 2 \min_{\overline{\Omega}} a(x),$$

then (1.10) is automatically satisfied. With suitable choices of the domain Ω and the coefficient $a(x)$, it is indeed possible to make Λ a single point. A couple of simple examples are in order; let $\Omega = B(0, R_0)$ be a two-dimensional disc of radius R_0 and $r = |x|$ for $x \in \Omega$. Let $v(x) = \xi_0(x)/a(x)$; (1.5) for ξ_0 may be rewritten, in the polar coordinate system, as

$$\begin{cases} v''(r) + v'(r)/r + v'(r)(\ln a(r))' + v(r)\Delta \ln a(r) = 1 \text{ in } (0, R_0), \\ v'(0) = 0, \quad v(R_0) = 0, \end{cases} \tag{1.11}$$

where $v(r)$ and $a(r)$ represent the functions v and a in the polar coordinates.

Example 1. If $a(r) = e^{-\frac{r^2}{4}}$, then $\Lambda = \{0\}$.

Since $\Delta \ln a(r) = -1$, $(\ln a(r))' = -r/2$, we have

$$v''(r) + (1/r - r/2)v'(r) - v(r) = 1 \text{ in } (0, R_0).$$

One may verify that $v(r) = z(r)/z(R_0) - 1$ is nonpositive in $[0, R_0]$, where

$$z(r) = r^2 + r^3 + \sum_{n=0}^{+\infty} \left(\frac{1}{4^{n+1}(n+2)!} r^{2(n+2)} + \frac{1}{2^{n+1}(2n+5)!!} r^{2n+5} \right).$$

$v(r)$ is a solution of problem (1.11). Since $z(r)$ is strictly increasing in $[0, R_0]$, so is $v(r)$. We know $|v(r)|$ takes its maximum value only at 0, that is, $\Lambda = \{0\}$.

Example 2. If $a(r) = 2(1 + r)$, then $\Lambda = \{0\}$.

In fact, let $\xi_0(r) = \frac{1}{3}r^3 + \frac{1}{2}r^2 - (\frac{1}{3}R_0^3 + \frac{1}{2}R_0^2)$. Then $\xi_0(r)$ is nonpositive in $[0, R_0]$, $\xi_0(R_0) = 0$, and ξ_0 is a solution of (1.11). $\xi_0 = \xi_0(r)$ is strictly increasing in $[0, R_0]$, so is $a = a(r)$. Therefore $|\xi_0(r)|/a(r)$ takes its maximum value only at $\{0\}$, so $\Lambda = \{0\}$.

For both of the above examples, depending on R_0, the thickness function a may take on values of different magnitude at different locations in the domain $B(0, R_0)$. It is interesting to note that the coefficient $a(x)$ takes its minimum value at the boundary in Example 1 but at the origin in Example 2. Based on the analysis given in this paper, near H_{c_1}, the solution of (1.3) with a single vortex in Ω will have its vortex pinned near the origin in both cases for small enough ε even though the origin is the thickest position in Example 1. This illustrates that the vortex pinning phenomenon may be affected by the competition between the applied field and the thickness variation.

We now state the second main theorem.

THEOREM 1.2. *For a solution sequence $u_n = u_{\varepsilon_n}$ of (1.3) given by the part (ii) of Theorem 1.1, up to a subsequence, there exist d points $c_i \in \Lambda$ such that $u_n \to u_*$ weakly in $W^{1,p}$ $(p < 2)$ and strongly in $H^1_{\text{loc}}(\Omega \setminus \cup_{i=1}^d \{c_i\})$, where u_* is a solution of*

$$\begin{cases} -\nabla \cdot (a(x)\nabla u_*) = a(x)u_*|\nabla u_*|^2 \text{ in } \Omega \setminus \cup_{i=1}^d \{c_i\}, \\ \frac{\partial u_*}{\partial \mathbf{n}} = 0 \qquad \text{on } \partial\Omega, \\ |u_*| = 1 \qquad \text{a.e. on } \Omega. \end{cases} \tag{1.12}$$

It is easy to see that the local minimizers in $\overline{D_M^a}$ may not be the solution of (1.3) (if it is on the boundary of $\overline{D_M^a}$). However, the vortex structure is only well defined for solutions that satisfy $|\nabla u| \le C/\varepsilon$. For this reason, similar to [23], we introduce a regularization as follows.

Let $u_\varepsilon^\gamma \in H^1(\Omega, \mathbb{R}^2)$ be a minimizer of the following minimization problem:

$$\min_{v \in H^1(\Omega, \mathbb{R}^2)} \left\{ \int_\Omega a(x) \left[\frac{1}{2}|\nabla v|^2 + \frac{1}{4\varepsilon^2}(1 - |v|^2)^2 \right] + \int_\Omega \frac{|v - u_\varepsilon|^2}{2\varepsilon^{2\gamma}} \right\}, \tag{1.13}$$

where $u_\varepsilon \in \overline{D_M^a}$. u_ε^γ is, in some sense, a regularization of u_ε in $\overline{D_M^a}$ and an a priori bound on the number of the vortices of u_ε^γ can be obtained. This in turn leads to a description of the vortices of u_ε. More careful examination of the minimizers u_ε of $J_a(u)$ in $\overline{D_M^a}$ shows that they are actually not on the boundary of $\overline{D_M^a}$, and hence they solve (1.3). For brevity, in the rest of the paper, unless explicitly stated to avoid ambiguity, the subscript ε is dropped from the notation u_ε and u_ε^γ; i.e., u and u^γ are used instead.

This paper is organized as follows. In the next section we shall give some basic estimates for $J_a(u)$ and for the regularization u^γ. The main ideas are to define the vortices of u^γ and to expand the energy $J_a(u)$. Using the idea of [23] and the estimate in [6], we may then give the lower bound for the energy. In section 3, we shall provide estimates to the critical magnetic field. In section 4, the proof Theorem 1.1 is given, and in section 5 we shall prove the convergence of the sequence of the minimizers, i.e., Theorem 1.2.

In the following discussion, we always consider the case $h_{ex} \le C|\ln\varepsilon|$ for some positive constant C and assume that the Abrikosov estimate $H_{c_1} \le C|\ln\varepsilon|$ holds.

2. Preliminaries. In this section we present technical estimates which can be proved by a slight modification of the results in [6, 23]. The detailed proofs are omitted. We begin by defining

$$J^0 = J_a(1) = \frac{1}{2}\int_\Omega \frac{1}{a(x)}|\nabla\xi|^2 \le Ch_{ex}^2. \tag{2.1}$$

LEMMA 2.1. *For $u \in \overline{D_M^a}$ minimizing $J_a(u)$ in $\overline{D_M^a}$, we have*

$$J_a(u) \le Ch_{ex}^2, \tag{2.2}$$

$$\int_\Omega a(x)|\nabla_{\mathbf{A}_0}u|^2 \le Ch_{ex}^2, \tag{2.3}$$

$$\frac{1}{4\varepsilon^2}\int_\Omega a(x)(1-|u|^2)^2 \le Ch_{ex}^2. \tag{2.4}$$

Proof. Taking $v \equiv 1$ as a comparison function leads to the results. ∎

For any $\tilde{u}$ with $J_a(\tilde{u}) \le Ch_{ex}^2$, let $\eta = |\tilde{u}|$. Since

$$a(x)|\nabla u - i\mathbf{A}_0 u|^2 = a(x)[|\nabla u|^2 + i\mathbf{A}_0(u^*\nabla u - u\nabla u^*) + |\mathbf{A}_0|^2|u|^2],$$

where u^* is the complex conjugate of u, we have the following lemma.

LEMMA 2.2. *For any $\tilde{u}$ with $J_a(\tilde{u}) \le Ch_{ex}^2$, we have*

$$J_a(\tilde{u}) = F_a(\tilde{u}) + \frac{1}{2}\int_\Omega \frac{1}{a(x)}|\nabla\xi|^2 + \int_\Omega (i\tilde{u}, \xi_{x_2}\tilde{u}_{x_1} - \xi_{x_1}\tilde{u}_{x_2}) + o(1).$$

LEMMA 2.3. *For $\tilde{u} \in \overline{D_M^a}$ such that $J_a(\tilde{u}) \le Ch_{ex}^2$, there exists $u \in \overline{D_M^a}$ such that*

$$|u| \le 1, \tag{2.5}$$

$$F_a(u) \le F_a(\tilde{u}), \tag{2.6}$$

$$J_a(u) \le J_a(\tilde{u}) + o(1). \tag{2.7}$$

If, in addition, $\tilde{u} \in \overline{D_M^a}$ is a minimizer of J_a in $\overline{D_M^a}$, then, as $\varepsilon \to 0$, there holds

$$F_a(u) = F_a(\tilde{u}) + o(1), \tag{2.8}$$

$$J_a(u) = J_a(\tilde{u}) + o(1). \tag{2.9}$$

LEMMA 2.4. *For $u \in \overline{D_M^a}$, we have $u^\gamma \in H^3(\Omega)$ (for any $0 < \gamma < 1$) which solves* (1.13) *and satisfies*

$$-\nabla\cdot(a(x)\nabla u^\gamma) = \frac{a(x)}{\varepsilon^2}u^\gamma(1-|u^\gamma|^2) + \frac{u-u^\gamma}{\varepsilon^{2\gamma}}, \tag{2.10}$$

$$F_a(u^\gamma) \le F_a(u) \le M|\ln\varepsilon|, \tag{2.11}$$

$$|u^\gamma| \le 1, |\nabla u^\gamma| \le \frac{C}{\varepsilon}. \tag{2.12}$$

This implies that $u^\gamma \in \overline{D_M^a}$. Taking u as a comparison function in (1.13) gives

$$\int_\Omega \frac{1}{2\varepsilon^{2\gamma}}|u-u^\gamma|^2 + F_a(u^\gamma) \le F_a(u) \le M|\ln\varepsilon|$$

so that $\|u-u^\gamma\|_{L^2(\Omega)} \le C\varepsilon^\gamma|\ln\varepsilon|^{\frac{1}{2}}$. Since $|\nabla u^\gamma| \le \frac{C}{\varepsilon}$, the vortices are well defined in the following sense.

LEMMA 2.5. *There exists $\lambda > 0$ and points a_i^ε $(i \in \mathcal{J}_1)$ in Ω with* $\mathrm{Card}\mathcal{J}_1 \le Ch_{ex}^2$ *such that*

$$|u^\gamma| \ge \frac{1}{2} \quad \text{in} \quad \Omega \setminus \cup_{i\in\mathcal{J}_1} B(a_i^\varepsilon, \lambda\varepsilon).$$

Proof. We know from [7] that there exists $\mu_0 > 0$ such that

$$\frac{1}{\varepsilon^2}\int_{B(a_i^\varepsilon,\lambda\varepsilon)} (1-|u^\gamma|^2)^2 \ge \mu_0 \quad \forall i \in \mathcal{J}_1.$$

Using exactly the same arguments given in [7], this implies that $\mathrm{Card}\mathcal{J}_1 \le Ch_{ex}^2$ since $J_a(u) \le Ch_{ex}^2$. □

The balls $B(a_i^\varepsilon, \lambda\varepsilon)$ are called "bad" discs and a_i^ε together with its degree d_i^ε is called a vortex of "size" $\lambda\varepsilon$. We now pay attention to the minimizer u^γ. Although a weight is added to the functional on u^γ, i.e., (1.13), the proofs of the following four lemmas on the properties of u^γ can still be obtained directly from the corresponding ones in [23] and [24] by replacing the energy density with $e_\varepsilon(u) = \frac{1}{2}a(x)[|\nabla u|^2 + \frac{1}{\varepsilon^2}(1-|u|^2)^2]$. We omit the details.

LEMMA 2.6. *For any $0 < \gamma < \beta < 1$, u^γ has no vortex (i.e., $|u^\gamma| \ge 1/2$) in* $\{x \in \Omega; \mathrm{dist}(x, \partial\Omega) \le \varepsilon^\beta\}$.

LEMMA 2.7. *For small enough ε,* $\mathrm{Card}\mathcal{J}_1$ *is uniformly bounded by a constant N which is independent of ε. Let $0 < \gamma < \beta < \mu < 1$ such that $\overline{\mu} = \mu^{N+1} > \beta$. For ε small enough, there exists a subset $\mathcal{J} \subset \mathcal{J}_1$ and a radius $\rho > 0$ with $\lambda\varepsilon \le \varepsilon^\mu \le \rho \le \varepsilon^{\overline{\mu}} < \varepsilon^\beta$ such that*

$$\begin{aligned}
&|u^\gamma| \ge 1/2 \quad \text{in } \Omega \setminus \cup_{i\in\mathcal{J}} B(a_i^\varepsilon, \rho),\\
&|u^\gamma| \ge 1 - 2|\ln\varepsilon|^{-2} \quad \text{on } \partial B(a_i^\varepsilon, \rho), \ \ i \in \mathcal{J},\\
&\int_{\partial B(a_i^\varepsilon,\rho)} e_\varepsilon(u^\gamma) \le C(\beta,\mu)/\rho, \quad i \in \mathcal{J},\\
&|a_i^\varepsilon - a_j^\varepsilon| \ge 8\rho, \quad i \ne j \in \mathcal{J}.
\end{aligned}$$

Denote $d_i^\varepsilon = \deg(u^\gamma, \partial B(a_i^\varepsilon, \rho))$. We have the following lemma.

LEMMA 2.8. *For small enough ε and $u \in \overline{D_M^a}$, $|d_i^\varepsilon| = O(1)$ for all $i \in \mathcal{J}$.*

Assume for the moment that $|\nabla u| \le C/\varepsilon$ which is true if u is shown to be a solution of (1.3); then, in the sense of [7], the vortices of u are well defined and there exists the same uniform bound on the vortex number. One may also have bigger vortices of size ρ (where "bigger" means $\rho \ge \lambda\varepsilon$), $(b_i^\varepsilon, q_i^\varepsilon)$, such that u satisfies the same conclusions as in Lemma 2.7 for u^γ. As in [23], we may compare $(a_i^\varepsilon, d_i^\varepsilon)$ (the vortices of u^γ) with $(b_i^\varepsilon, q_i^\varepsilon)$ (the vortices of u) by the minimal connection between the vortices.

LEMMA 2.9. *For small ε, there holds* $\mathrm{dist}(a,b) \le C\varepsilon^\gamma|\ln\varepsilon|$

For the definition of $\mathrm{dist}(a,b)$ and the proof of this lemma, we refer to [23]. The following lemma gives the splitting of the energy $J_a(u)$ as in [23].

LEMMA 2.10. *For any $\tilde{u}$ satisfying* (2.2)–(2.4)*, let u be associated to $\tilde{u}$ as in Lemma* 2.3 *and u^γ be associated to u by solving the minimization problem* (1.13) *with vortices (a_i, d_i) satisfying Lemma* 2.7*. Then we have*

$$J_a(u) = F_a(u) + \frac{1}{2}\int_\Omega \frac{1}{a(x)}|\nabla\xi|^2 + 2\pi\sum_{i\in\mathcal{J}} d_i\xi(a_i), \quad as\ \varepsilon\to 0,$$

where $\xi = h_{ex}\xi_0$ and ξ_0 is the unique solution of problem (1.5).

Using this splitting, we have the following lemma.

LEMMA 2.11. *The constant J^0 in* (2.1) *is asymptotically equal to the minimal energy among vortexless configurations; i.e.,* $\inf_{\{u:\mathcal{J}=\emptyset\}} J_a(u) = J^0 + o(1)$ *as $\varepsilon\to 0$.*

Let $e_\varepsilon(u) = \frac{1}{2}a(x)[|\nabla u|^2 + \frac{1}{2\varepsilon^2}(1-|u|^2)^2]$ and $\Omega_\rho = \Omega\setminus\cup_{i\in\mathcal{J}}B(a_i,\rho)$, where $B(a_i,\rho)$'s are defined in Lemma 2.7. We have the following lemma.

LEMMA 2.12. *Assume that $\mathcal{J} = \{1, 2, \dots, k\}$; then*

$$\frac{1}{2}\int_{\Omega_\rho} a(x)|\nabla u^\gamma|^2 \ge \pi\sum_{i\in\mathcal{J}} a(a_i)d_i^2|\ln\rho| + W((a_1,d_1),\dots,(a_k,d_k)) + O(1),$$

where

$$W((a_1,d_1),\dots,(a_k,d_k)) = -\pi\sum_{i\ne j\in\mathcal{J}} a(a_i)d_id_j\ln|a_i-a_j| - \pi\sum_{i\in\mathcal{J}} d_iR_0(a_i)$$

and $R_0(x) = \Phi_0(x) - \sum_{i\in\mathcal{J}} a(a_i)d_i\ln|x-a_i|$ with $\Phi_0(x)$ solves

$$\begin{cases} -\mathrm{div}(\frac{1}{a(x)}\nabla\Phi_0) = 2\pi\sum_{i\in\mathcal{J}} d_i\delta_{a_i} & in\ \Omega,\\ \Phi_0 = 0 \qquad on\ \partial\Omega. \end{cases}$$

In the following lemma, we give a few more precise lower bounds on $F_a(u^\gamma)$.

LEMMA 2.13. *For ε and ρ satisfying Lemma* 2.7*, we have*

$$F_a(u^\gamma) \ge \pi\sum_{i\in\mathcal{J}} a(a_i)[d_i^2|\ln\rho| + |d_i||\ln(\rho/\varepsilon)| + W((a_1,d_1),\dots,(a_k,d_k)) + O(1), \tag{2.13}$$

$$F_a(u^\gamma) \ge \pi\sum_{i\in\mathcal{J}} a(a_i)|d_i||\ln(\rho/\varepsilon)| + O(1). \tag{2.14}$$

3. Obtaining the critical magnetic field H_{c_1}. Using the splitting and the lower bound of $J_a(u)$, we now estimate the critical magnetic field H_{c_1}.

LEMMA 3.1. *Let $h_{ex} = k_a|\ln\varepsilon| + o(|\ln\varepsilon|)$ and $\tilde{u}\in\overline{D_M^a}$ be a minimizer of $J_a(u)$ in $\overline{D_M^a}$ and $\{(a_i,d_i) : i\in\mathcal{J}\}$ be the vortices of u^γ. For ε small enough, if $\mathcal{J}\ne\emptyset$, say, $\mathcal{J} = \{1,\dots,k\}$, then*

(i) *$d_i > 0$ for any $i\in\mathcal{J}$, and*

(ii) *$\mathrm{dist}(a_i,\partial\Omega)\ge\alpha>0$ for some positive constant α, and consequently*

(iii) *$W((a_1,d_1),\dots,(a_k,d_k))\ge C$ for some constant C.*

Proof. We divide the proof into two steps.

Step 1. We first prove that, for ε small enough, $d_i>0$ for $i\in\mathcal{J}$. Since $\tilde{u}\in\overline{D_M^a}$ is a minimizer of $J_a(u)$, it follows from Lemma 2.11 that $J_a(u)\le J^0+o(1)$, i.e.,

$$F_a(u) + J^0 + 2\pi h_{ex}\sum_{i\in\mathcal{J}} d_i\xi_0(a_i) + o(1) \le J^0 + o(1).$$

Therefore

$$F_a(u) \le -2\pi h_{ex} \sum_{i\in\mathcal{J}} d_i \xi_0(a_i) + o(1) \tag{3.1}$$

or equivalently (noting that $\xi_0 < 0$ in Ω)

$$\begin{aligned} F_a(u) &\le 2\pi(k_a|\ln\varepsilon| + o(|\ln\varepsilon|)) \max\left|\frac{\xi_0(x)}{a(x)}\right| \sum_{d_i>0} a(a_i)d_i + o(1) \\ &\le \pi|\ln\varepsilon| \sum_{d_i>0} a(a_i)d_i + o(|\ln\varepsilon|). \end{aligned}$$

This inequality implies

$$F_a(u^\gamma) \le F_a(u) \le \pi|\ln\varepsilon| \sum_{d_i>0} a(a_i)d_i + o(|\ln\varepsilon|). \tag{3.2}$$

Combining (3.2) with (2.14) in Lemma 2.13 we obtain

$$\pi(1-\mu)\left(\sum_{i\in\mathcal{J}} a(a_i)|d_i|\right)|\ln\varepsilon| \le \pi\left(\sum_{d_i>0} a(a_i)d_i\right)|\ln\varepsilon| + o(|\ln\varepsilon|) \tag{3.3}$$

since $\varepsilon^\mu \le \rho \le \varepsilon^{\overline{\mu}}$. This implies

$$(1-\mu)\sum_{d_i<0} a(a_i)|d_i| \le \mu \sum_{d_i>0} a(a_i)d_i + o(1). \tag{3.4}$$

We estimate the first term on the right-hand side of (3.4). By (2.11), (2.14),

$$\mu \sum_{d_i>0} a(a_i)d_i \le \mu \sum_{i\in\mathcal{J}} a(a_i)|d_i| \le M\mu/(\pi(1-\mu)) + o(1).$$

Substituting this into (3.4), we get

$$\sum_{d_i<0} a(a_i)|d_i| \le M\mu/(\pi(1-\mu)^2) + o(1).$$

This means $\{i \in \mathcal{J};\ d_i < 0\} = \emptyset$ if one chooses μ small enough.

Step 2. We prove (ii) and (iii) in this step. It follows from Step 1 that

$$-\pi \sum_{i\ne j} a(a_i)d_i d_j \ln|a_i - a_j| \ge O(1)$$

and then

$$W((a_1,d_1),\dots,(a_k,d_k)) \ge -\pi \sum_{i\in\mathcal{J}} d_i R_0(a_i) + O(1). \tag{3.5}$$

For the proof of $\|R_0\|_{L^\infty(\Omega)} \le C$, similar to [23] and [6], it suffices to prove that $\operatorname{dist}(a_i, \partial\Omega)$ is uniformly bounded from below. Indeed, it can be shown as in [23] that

$$\|R_0(x)\|_{L^\infty(\Omega)} \le C\beta|\ln\varepsilon| + O(1). \tag{3.6}$$

Therefore we deduce

$$W((a_1,d_1),\dots,(a_k,d_k)) \geq -C\beta|\ln\varepsilon|. \tag{3.7}$$

On the other hand, we know from (3.2) and (2.13) (in view of $d_i^2 \geq d_i > 0$) that

$$\begin{aligned} F_a(u^\gamma) &\leq F_a(u) \leq -2\pi h_{ex}\sum_{i\in\mathcal{J}} d_i\xi_0(a_i) + o(1),\\ F_a(u^\gamma) &\geq \pi\sum_{i\in\mathcal{J}} a(a_i)d_i|\ln\varepsilon| + W((a_1,d_1),\dots,(a_k,d_k)) + O(1)\\ &\geq \pi\sum_{i\in\mathcal{J}} a(a_i)d_i|\ln\varepsilon| - C\beta|\ln\varepsilon| + O(1). \end{aligned}$$

Putting these two inequalities together and using

$$h_{ex} = k_a|\ln\varepsilon| + o(|\ln\varepsilon|) = \frac{|\ln\varepsilon|}{2\max_\Omega|\xi_0(x)/a(x)|} + o(|\ln\varepsilon|),$$

we get

$$2\pi h_{ex}\sum_{i\in\mathcal{J}} a(a_i)d_i\left[\frac{\xi_0(a_i)}{a(a_i)} + \max_\Omega\left|\frac{\xi_0(x)}{a(x)}\right|\right] \leq C\beta|\ln\varepsilon| + o(|\ln\varepsilon|). \tag{3.8}$$

Since $d_i \geq 1$ and $a(a_i) \geq \alpha_0 > 0$ for $i \in \mathcal{J}$, the above implies

$$\frac{\xi_0(a_i)}{a(a_i)} + \max_\Omega\left|\frac{\xi_0(x)}{a(x)}\right| \leq C\beta\max_\Omega\left|\frac{\xi_0(x)}{a(x)}\right| \ \ \forall i\in\mathcal{J}. \tag{3.9}$$

Taking $\beta > 0$ such that $C\beta < 1/2$, we get

$$\frac{\xi_0(a_i)}{a(a_i)} \leq -\frac{1}{2}\max_\Omega\left|\frac{\xi_0(x)}{a(x)}\right| < 0. \tag{3.10}$$

Since $\xi_0 = 0$ on $\partial\Omega$, we thus have $\operatorname{dist}(a_i,\partial\Omega)$ being uniformly bounded from below. So, $\|R_0\|_{L^\infty(\Omega)} \leq C$. This implies $W \geq O(1)$ uniformly by (3.5). □

Now, let $\overline{D_0} = \{u \in \overline{D_M^a}; \mathcal{J} = \emptyset\}$, and we have the following lemma.

LEMMA 3.2. *Suppose* $\max_{\overline{\Omega}} a(x)\pi < M$. *There are* $k_2^\varepsilon = O(1)$, $k_3^\varepsilon = o(1)$, *and* $\varepsilon_0 > 0$ *such that, for* $h_{ex} = |\ln\varepsilon|/(2\max_\Omega|\xi_0(x)/a(x)|) + t$, *there holds*

(i) *if* $t < k_2^\varepsilon$ *and* $\tilde{u}$ *is a minimizer of* J_a *in* $\overline{D_M^a}$, *then* $\mathcal{J} = \emptyset$ *and*

$$J_a(\tilde{u}) = \inf_{\overline{D_0}} J_a(u) = J^0 + o(1);$$

(ii) *if* $t = k_2^\varepsilon$, *there is* $u \in \overline{D_M^a}$ *with a simple vortex and* $J_a(u) \leq \inf_{\overline{D_0}} J_a(v)$;

(iii) *if* $t \geq k_2^\varepsilon + k_3^\varepsilon$, *there is* $u \in \overline{D_M^a}$ *with a simple vortex and* $J_a(u) < \inf_{\overline{D_0}} J_a(v)$.

Proof. Let J^0 be as in (2.1). We have

$$J_a(u) = F_a(u) + J^0 + 2\pi h_{ex}\sum_{i\in\mathcal{J}} d_i\xi_0(a_i) + o(1).$$

Clearly, $J^0 = \inf_{\overline{D_0}} J_a(v)$. If $\mathcal{J} \neq \emptyset$, then we consider two cases.

Case 1. $h_{ex} \le (1-\mu^*)k_a|\ln\varepsilon|$ for some $0<\mu^*<1$. Since $\rho \ge \varepsilon^\mu$ for some $\mu>0$, it follows from Lemma 2.13 that

$$F_a(u) \ge F_a(u^\gamma) \ge (1-\mu)\pi \sum_{i\in\mathcal{J}} a(a_i)|d_i||\ln\varepsilon| + O(1), \tag{3.11}$$

and then

$$J_a(u) \ge J^0 + \pi(1-\mu)\sum_{i\in\mathcal{J}} a(a_i)|d_i||\ln\varepsilon| - 2\pi h_{ex}\sum_{i\in\mathcal{J}} a(a_i)|d_i|\left|\frac{\xi_0(a_i)}{a(a_i)}\right| + O(1).$$

Hence, $J_a(u) > \inf_{\overline{D_0}} J_a(v)$ as long as

$$2\pi h_{ex}\sum_{i\in\mathcal{J}} a(a_i)|d_i|\left|\frac{\xi_0(a_i)}{a(a_i)}\right| \le (1-\mu)\pi\sum_{i\in\mathcal{J}} a(a_i)|d_i||\ln\varepsilon| + O(1)$$

which may be valid if we take $\mu<\mu^*$ since

$$h_{ex} \le (1-\mu)\frac{|\ln\varepsilon|}{2\max\limits_{\Omega}|\xi_0(x)/a(x)|}.$$

Case 2. $t<k_2^\varepsilon$ with $|t| = o(|\ln\varepsilon|)$. Then, by Lemma 3.1, we have

$$W((a_1,d_1),\dots,(a_k,d_k)) \ge C$$

for some constant C, thus, by Lemma 2.13, we get

$$F_a(u) \ge F_a(u^\gamma) \ge \pi\sum_{i\in\mathcal{J}} a(a_i)|d_i||\ln\varepsilon| + O(1).$$

Then, similar to Case 1, we have $J_a(u) > \inf_{\overline{D_0}} J_a(v)$ as long as

$$h_{ex} \le \frac{|\ln\varepsilon|}{2\max\limits_{\Omega}|\xi_0(x)/a(x)|} + O(1).$$

This verifies conclusion (i) in the lemma.

Next, let $k_a = 1/(2\max_\Omega|\xi_0(x)/a(x)|)$. As in [23], set

$$Z^\varepsilon = \Big\{ t\in\mathbb{R};\ \text{there exists } u\in\overline{D_M^a} \text{ with at least one vortex}$$
$$\text{and } J_a(u) < \inf_{\{\mathcal{J}=\emptyset\}} J_a \text{ for } h_{ex} = k_a|\ln\varepsilon| + t\Big\}.$$

In the following, we prove $Z^\varepsilon\neq\emptyset$ which would allow us to define $k_2^\varepsilon = \inf Z^\varepsilon$ and to prove that there exists $k_3^\varepsilon = o(1)$ such that $[k_2^\varepsilon + k_3^\varepsilon, +\infty] \subset Z^\varepsilon$.

Let $c\in\Omega$ such that $|\xi_0(c)/a(c)| = \max_\Omega|\xi_0(x)/a(x)|$. Consider the problem

$$\nu_\varepsilon(c) = \min_W \frac{1}{2}\int_{\Omega\setminus B(c,\varepsilon)} a(x)|\nabla u|^2, \tag{3.12}$$

where $W = \{\, u\in H^1(\Omega\setminus B(c,\varepsilon), S^1), \deg(u,\partial B(c,\varepsilon)) = 1\,\}$. Similar as before,

$$\nu_\varepsilon(c) = \pi a(c)|\ln\varepsilon| + O(1).$$

Let u be a minimizer of problem (3.12) which is well defined on $\Omega\setminus B(c,\varepsilon)$. Extending u to the whole domain Ω by defining it on $B(c,\varepsilon)$ as in [23] and denoting it by $\bar{u}$, we may get, as similarly done in [23],

$$F_a(\overline{u},\Omega)=F_a(\overline{u},B(c,\varepsilon))+\frac{1}{2}\int_{\Omega\setminus B(c,\varepsilon)}a(x)|\nabla u|^2\le K+a(c)\pi|\ln\varepsilon|.$$

For $h_{ex}=\frac{1}{2\max|\xi_0(x)/a(x)|}|\ln\varepsilon|+t=-\frac{1}{2\xi_0(c)/a(c)}|\ln\varepsilon|+t$, we have

$$\begin{aligned}J_a(\overline{u})&\le F_a(\overline{u})+J^0+2\pi h_{ex}\xi_0(c)+o(1)\\&=K-2\pi|\xi_0(c)|t+J^0+o(1).\end{aligned}$$

This implies $t\in Z^\varepsilon$ when $2\pi|\xi_0(c)|t\ge K+o(1)$. So, $Z^\varepsilon\neq\emptyset$ and $k_2^\varepsilon=\inf Z^\varepsilon\le K/2\pi|\xi_0(c)|+o(1)$. On the other hand, $h_{ex}\le k_a|\ln\varepsilon|+O(1)$; we thus know $k_2^\varepsilon\ge O(1)$ which gives $k_2^\varepsilon=O(1)$.

Finally, we prove that there exists $k_3^\varepsilon=o(1)$ such that $[k_2^\varepsilon+k_3^\varepsilon,+\infty]\subset Z^\varepsilon$. In fact, let $t\in Z^\varepsilon$ and, for $h_{ex,1}=k_a|\ln\varepsilon|+t$, $J_a(u)<\inf_{\overline{D_0}}J_a(v)$ and u^γ has vortices (a_i,d_i). Assume $t^{'}>t$ and $h_{ex,2}=k_a|\ln\varepsilon|+t^{'}$; we have

$$\begin{aligned}J_a(u)&=F_a(u)+J^0+2\pi h_{ex,2}\sum_{i=1}^k\xi_0(a_i)d_i+o(1)\\&\le J_a(u)+o(1)-(t^{'}-t)\sum_{i=1}^k2\pi|\xi_0(a_i)|d_i.\end{aligned}$$

Thus, if $t^{'}-t\ge k_3^\varepsilon=o(1)$, then $J_a(u)<\inf_{\{\mathcal{J}=\emptyset\}}J_a$, i.e., $[k_2^\varepsilon+k_3^\varepsilon,+\infty]\subset Z^\varepsilon$.

In summary, we have deduced that $H_{c_1}=k_a|\ln\varepsilon|+k_2^\varepsilon$ for the lower critical field. This completes the proof of the lemma. □

4. Proof of Theorem 1.1. By Lemma 3.2 and the bounds in Lemma 2.3 and Lemma 2.11, we see that, for $h_{ex}\le H_{c_1}$, the minimizer of J_a in $\overline{D_M^a}$ has no vortex and it is in the interior of $\overline{D_M^a}$, thus the first part of Theorem 1.1 follows.

To complete the proof of Theorem 1.1, we need the following lemmas.

LEMMA 4.1. *Let $h_{ex}=k_a|\ln\varepsilon|+o(|\ln\varepsilon|)$ and $\tilde{u}\in\overline{D_M^a}$ be a minimizer of $J_a(u)$ in $\overline{D_M^a}$. $\{(a_i,d_i)\}_{i=1}^k$ are the vortices of u^γ. Then $d_i=1$ for any $i\in\mathcal{J}=\{1,\dots,k\}$.*

Proof. Using the lower bound on W proved in Lemma 3.1 and returning to Lemma 2.13, we have

$$\pi\sum_{i\in\mathcal{J}}a(a_i)d_i^2|\ln\rho|+\pi\sum_{i\in\mathcal{J}}a(a_i)d_i\ln\frac{\rho}{\varepsilon}+O(1)\le\pi\sum_{d_i>0}a(a_i)d_i|\ln\varepsilon|+o(|\ln\varepsilon|).$$

This gives

$$\pi\sum_{i\in\mathcal{J}}a(a_i)(d_i^2-d_i)|\ln\rho|\le o(|\ln\varepsilon|).$$

Therefore we have from $\rho\ge\varepsilon^\mu$ that

$$\mu\sum_{i\in\mathcal{J}}a(a_i)(d_i^2-d_i)|\ln\varepsilon|\le\sum_{i\in\mathcal{J}}a(a_i)(d_i^2-d_i)|\ln\rho|\le o(|\ln\varepsilon|),$$

which implies

$$\mu \min_{\Omega} a(x) \sum_{i\in\mathcal{J}} (d_i^2 - d_i) \le o(1).$$

This inequality is impossible if there is a $d_i > 1$ for small ε. So when $\varepsilon \le \varepsilon_0$, we have $d_i = 1$ for all $i \in \mathcal{J}$. The lemma is proved. □

We are now closer to a complete proof of Theorem 1.1. Consider

$$H_{c_1} + k_3^\varepsilon \le h_{ex} \le k_a |\ln\varepsilon| + O(1),$$

where

$$H_{c_1} = k_a |\ln\varepsilon| + k_2^\varepsilon, \qquad k_a = 1/(2\max_{\Omega} |\xi_0(x)/a(x)|),$$

$k_2^\varepsilon = O(1)$, and $k_3^\varepsilon = o(1)$. Let

$$\Lambda = \left\{ x, x \in \Omega : \left| \frac{\xi_0(x)}{a(x)} \right| = \max_{y\in\Omega} \left| \frac{\xi_0(y)}{a(y)} \right| \right\}.$$

The proof of Theorem 1.1 can be obtained by proving the following three lemmas.

LEMMA 4.2. *Let $u \in \overline{D_M^a}$ be a minimizer of $J_a(u)$, and let (a_i, d_i) ($d_i = 1$ for all $i \in \mathcal{J}$) be the vortices of u^γ. Then*

$$\operatorname{dist}(a_i, \Lambda) \to 0, \text{ as } \varepsilon \to 0 \ \forall i \in \mathcal{J}, \tag{4.1}$$

$$\operatorname{dist}(a_i, a_j) \ge \alpha > 0 \ \forall i \ne j \in \mathcal{J}. \tag{4.2}$$

The first result is also true under the assumption $h_{ex} \le k_a |\ln\varepsilon| + o(|\ln\varepsilon|)$.

Proof. If $\mathcal{J} \ne \emptyset$, we have from Lemma 4.1 that $d_i = 1$ for all $i \in \mathcal{J}$. Now let $d = \sum_{i\in\mathcal{J}} d_i = \operatorname{Card}\mathcal{J} = \deg(u^\gamma, \partial\Omega)$. It follows from Step 1 in the proof of Lemma 3.1 and from Lemma 2.13 that

$$W(a_1, \dots, a_k) + \pi \sum_{i\in\mathcal{J}} a(a_i) |\ln\varepsilon| + O(1) \le F_a(u^\gamma) \le -2\pi h_{ex} \sum_{i\in\mathcal{J}} \xi_0(a_i) + o(|\ln\varepsilon|).$$

Then

$$2\pi h_{ex} \sum_{i\in\mathcal{J}} a(a_i) \left(\frac{\xi_0(a_i)}{a(a_i)} + \max_{\Omega} \left| \frac{\xi_0(x)}{a(x)} \right| \right) \le o(|\ln\varepsilon|).$$

Hence we have

$$\sum_{i\in\mathcal{J}} \left(\frac{\xi_0(a_i)}{a(a_i)} + \max_{\Omega} \left| \frac{\xi_0(x)}{a(x)} \right| \right) \le \frac{o(|\ln\varepsilon|)}{h_{ex}} \to 0.$$

This implies the first conclusion

$$\operatorname{dist}(a_i, \Lambda) \to 0, \text{ as } \varepsilon \to 0 \ \forall i \in \mathcal{J}.$$

Moreover, since $W(a_1, \dots, a_k) \ge O(1)$ and

$$W(a_1, \dots, a_k) + \pi \sum_{i\in\mathcal{J}} a(a_i) |\ln\varepsilon| + O(1) \le -2\pi h_{ex} \pi \sum_{i\in\mathcal{J}} \xi_0(a_i) + O(1),$$

we have $W(a_1,\dots,a_k)\le O(1)$ if $h_{ex}\le k_a|\ln\varepsilon|+O(1)$. Therefore we conclude that $|a_i-a_j|$ remains bounded from below uniformly, since as in [7] we could prove that $W\to+\infty$ if $|a_i-a_j|\to 0$ for some $i\neq j$. Lemma 4.2 is proved. □

LEMMA 4.3. *Let M, n satisfy Assumption* 1.1*, and let $\tilde u$ be a minimizer of $J_a(u)$ in $\overline{D_M^a}$; then $\tilde u$ satisfies* (1.3) *and $u=\tilde u$, where u is defined by $\tilde u$ as in Lemma* 2.3.

Proof. It suffices to prove that $\tilde u$ is not on the boundary of $\overline{D_M^a}$. Since we have proved that W is a bounded quantity and $\operatorname{dist}(a_i,a_j)\ge\alpha>0$, we get

$$\begin{aligned}\pi\sum_{i\in\mathcal{J}}a(a_i)|\ln\varepsilon|+O(1)&\le -2\pi h_{ex}\sum_{i\in\mathcal{J}}\xi_0(a_i)+O(1)\\&=2\pi h_{ex}\sum_{i\in\mathcal{J}}a(a_i)\,|\xi_0(a_i)/a(a_i)|+O(1)\\&\le\pi\sum_{i\in\mathcal{J}}a(a_i)|\ln\varepsilon|+O(1).\end{aligned}$$

This inequality and Lemma 2.3 yield that

$$F_a(u)=F_a(\tilde u)+o(1)=\pi\sum_{i\in\mathcal{J}}a(a_i)|\ln\varepsilon|+O(1)\le M|\ln\varepsilon|+O(1).$$

So, $\sum_{i\in\mathcal{J}}a(a_i)\le M/\pi$. It follows from Lemma 4.2 that as $\varepsilon\to0$, $a_i\to c_i\in\Lambda$. Then, for ε small enough, $d\le M/(\pi m_d)$, where $m_d=(\sum_{i\in\mathcal{J}}a(c_i))/d$ and

$$\frac{M}{\pi m_d}\in\left[\frac{M}{\pi\max\limits_{\Lambda}a(x)},\quad\frac{M}{\pi\min\limits_{\Lambda}a(x)}\right]\subset(n,n+1).$$

Thus, $M/(\pi m_d)$ is not an integer which implies $d<M/(\pi m_d)$. Hence $\pi\sum_{i\in\mathcal{J}}a(a_i)<M$ for $\varepsilon\le\varepsilon_0$. Thus, there is a positive number $\eta>0$ such that

$$F_a(\tilde u)\le\pi\sum_{i\in\mathcal{J}}a(a_i)|\ln\varepsilon|+O(1)\le(M-\eta)|\ln\varepsilon|,$$

which means that $\tilde u$ is not on $\partial\overline{D_M^a}$. The lemma is proved. □

REMARK 4.1. *It follows from the proof of Lemma* 4.3 *that $d<M/(\pi\max_\Lambda a(x))$. Otherwise, since we also have $d\min_\Lambda a(x)\le\sum_{i=1}^d a(a_i)\le M/\pi$, this implies that*

$$d\in\left[\frac{M}{\pi\max\limits_{\Lambda}a(x)},\quad\frac{M}{\pi\min\limits_{\Lambda}a(x)}\right]\subset(n,n+1);$$

then d is not an integer. This leads to a contradiction.

Now we may continue the proof of Theorem 1.1 with the following lemma. Once u is a solution of (1.3), we may show that $|\nabla u|\le C/\varepsilon$. Then u has bigger vortices of size ρ: $\{(b_i,q_i)\}_{i\in\mathcal{J}}$. The following lemma compares what we call the bigger vortices of u (i.e., the vortices of u^γ) with the real vortices of u. Its proof follows easily from the same arguments given in [23].

LEMMA 4.4. *For sufficiently small ε, we have*

(i) *if u is a solution of* (1.3) *such that $J_a(u)\le Ch_{ex}^2$, then $|u|\le1$ and there exists a constant $C>0$ such that $|\nabla u|\le C/\varepsilon$;*

(ii) *if u is a solution of* (1.3) *such that u^γ has no vortices (i.e., $|u^\gamma| \geq 1/2$) and $J_a(u) \leq J^0$, then u has no vortices on Ω ($|u| \geq 1/2$);*

(iii) *if u is a solution given in Theorem* 1.1*, then its vortices (of size ρ) satisfy the same conclusions as those of u^γ;*

(iv) *if, in addition, $\{a_i\}_{i\in\mathcal{J}}$ are the vortices of u^γ of degree one, then the vortices $\{b_i\}_{i\in\mathcal{J}}$ are also of degree one and Lemma* 2.7 *(on u^γ) is satisfied by u.*

The following lemma shows that the real vortices of u remain far from the boundary.

LEMMA 4.5. *If $u \in D_M^a$ is an energy minimizer satisfying* (1.3)*, then, for any $0 < \beta < 1$, $|u| \geq 1/2$ on $\{x \in \Omega : \mathrm{dist}(x, \partial\Omega) \leq \varepsilon^\beta\}$. Moreover, u has no zero degree vortex.*

Finally, since $\mathrm{dist}(a_i, \partial\Omega)$, $\mathrm{dist}(b_i, \partial\Omega)$ remain bounded from below by a positive constant and $\mathrm{dist}(a, b) \leq C\varepsilon^\gamma |\ln\varepsilon|$, we have for small ε that $\mathbb{R}^2$ is a hole of null multiplicity. This implies $\sum_i q_i = \sum_i q_i = \mathrm{Card}\mathcal{J}'$, and the b_i's tend to the a_i's with the same multiplicities. However, $\inf_{i\neq j} |a_i - a_j| \leq C|\ln\varepsilon|^{-\frac{1}{2}}$; comparing with $\mathrm{dist}(a, b) \leq C\varepsilon^\gamma |\ln\varepsilon|$, the b_i's must be of multiplicity one, and $q_i = 1$ for all $i \in \mathcal{J}'$. Theorem 1.1 is proved.

5. Proof of Theorem 1.2. In this section we derive the convergence of u_ε and the limit equation of (1.3). The case $d = 0$ is again trivial to consider; we omit the details. Now, for $H_{c_1} + k_3^\varepsilon \leq h_{ex} \leq H_{c_1} + O(1)$, let us consider a sequence $\varepsilon_n \to 0$ and denote $u_n = u_{\varepsilon_n}$ an associated solution of (1.3) given by Theorem 1.1.

We also denote $\{b_i\}_{i\in\mathcal{J}}$ the real vortices of u_n (see Lemma 4.4) of size $\lambda\varepsilon$, and $\{b_i\}_{i\in\mathcal{J}'\subset\mathcal{J}}$ its vortices of size ρ, exactly as we did for u^γ. (Again, the superscript ε in the notation of b_i is removed.) This means

$$|u_n| \geq 1/2 \ \text{ on } \Omega \setminus \cup_{i\in\mathcal{J}} B(b_i, \lambda\varepsilon) \ \text{ and on } \Omega \setminus \cup_{i\in\mathcal{J}'} B(b_i, \rho),$$
$$\cup_{i\in\mathcal{J}} B(b_i, \lambda\varepsilon) \subset \cup_{i\in\mathcal{J}'} B(b_i, \rho).$$

Extracting a subsequence if necessary, we may assume that $\mathrm{Card}\mathcal{J}' \equiv d \geq 1$ and

$$b_i^{\varepsilon_n} \to c_i \in \Lambda \ \text{ for } i \in \mathcal{J}'.$$

Here, we put back the superscript ε_n to avoid ambiguity. First, we prove that

$$|\ln\varepsilon_n|\nabla u_n \to 0, \quad \text{strongly in} \quad L^p(\Omega)\ \forall p < 2.$$

In fact, we may rewrite (1.3) as

$$-\nabla\cdot(a(x)\nabla u) + 2ia(x)A_0\cdot\nabla u = a(x)u\left[|A_0|^2 + \frac{1}{\varepsilon^2}(1 - |u|^2)\right]. \tag{5.1}$$

This equation is equivalent to the following system if we write locally $u = \rho e^{i\varphi}$:

$$\begin{cases} -\nabla\cdot(a(x)\nabla\rho) + a(x)\rho|\nabla\varphi|^2 - 2a\rho A_0\cdot\nabla\varphi = a(x)\rho(|A_0|^2 + \frac{1-\rho^2}{\varepsilon^2}), \\ -\nabla\cdot(a(x)\rho^2\nabla\varphi) + a(x)A_0\cdot\nabla\rho^2 = 0. \end{cases} \tag{5.2}$$

Define

$$\overline{\rho} = \max\{\rho, 1 - |\ln\varepsilon|^{-4}\} \geq 1 - |\ln\varepsilon|^{-4}, \quad K = \{x \in \Omega, \rho \geq 1 - |\ln\varepsilon|^{-4}\};$$

then $\nabla\overline{\rho} = \nabla\rho$ on K, and $\nabla\overline{\rho} = 0$ on $\Omega \setminus K$. It follows from

$$\frac{1}{\varepsilon^2}\int_\Omega (1-\rho)^2 \le \frac{1}{\varepsilon^2}\int_\Omega (1-\rho^2)^2 \le C|\ln\varepsilon|^2$$

that $\mathrm{meas}(\Omega \setminus K) \le C\varepsilon^2|\ln\varepsilon|^{10}$.

Multiplying the first equation in (5.2) by $1-\overline{\rho}$, and integrating over Ω, we get

$$2\int_K a(x)\rho|\nabla\rho|^2 \le C\int_\Omega a(x)(1-\overline{\rho})\rho[|A_0|^2 + |A_0||\nabla u|] \le C\|1-\overline{\rho}\|_{L^\infty}|\ln\varepsilon|^2.$$

This inequality, together with the fact $0 \le 1-\overline{\rho} \le |\ln\varepsilon|^{-4}$, yields

$$\int_K |\nabla\rho|^2 \le C/|\ln\varepsilon|^2 \to 0.$$

On the other hand, we have for $p < 2$

$$\begin{aligned}\int_{\Omega\setminus K}|\nabla\rho|^p &\le \left(\int_{\Omega\setminus K}|\nabla\rho|^2\right)^{p/2}\mathrm{meas}(\Omega\setminus \mathrm{K})^{1-p/2}\\ &\le C\varepsilon^{2-p}|\ln\varepsilon|^{10-4p} \le C/|\ln\varepsilon|^2 \to 0.\end{aligned}$$

Combining the above two estimates, we have

$$\int_\Omega |h_{ex}\nabla\rho|^p \le \frac{C}{|\ln\varepsilon|^{2-p}} \to 0. \tag{5.3}$$

Now we rewrite the second equation of (5.2) as

$$-\nabla\cdot(a(x)\rho^2\nabla\varphi) = f(x),$$

where $f(x) = -a(x)A_0\cdot\nabla\rho^2$. Since $a(x)A_0 = h_{ex}\nabla^\perp\xi_0$ is a smooth function, we have

$$\int_\Omega |f(x)|^p \le C/|\ln\varepsilon|^{2-p}.$$

Hence we have

$$\int_\Omega |h_{ex}\nabla\varphi|^p \le C/|\ln\varepsilon|^{2-p} \to 0. \tag{5.4}$$

The estimates (5.3) and (5.4) yield

$$\int_\Omega |h_{ex}\nabla u|^p \to 0.$$

It follows that, for any smooth test function $\phi \in C_0^\infty(\Omega)$,

$$2i\int_\Omega a\,\phi A_0\cdot\nabla u = 2i\int_\Omega(\phi\nabla^\perp\xi_0)\cdot(h_{ex}\nabla u) \to 0 \tag{5.5}$$

since we may take $(\phi\nabla^\perp\xi_0) \in L^q(\Omega)$ for any $q > 2$.

Let s be a positive integer. As in the proofs of Lemmas 2.12 and 2.13, by solving an auxiliary problem, we may get

$$\int_{B(c_i,\frac{1}{s})\setminus B(b_i,\rho)} a(x)|\nabla u_n|^2 \geq \pi a(b_i)|\ln\rho| + O(1) \qquad \forall i \in \mathcal{J}'$$

and

$$\int_{B(b_i,\rho)} e_\varepsilon(u_n) \geq \pi a(b_i)\ln\frac{\rho}{\varepsilon} + O(1) \qquad \forall i \in \mathcal{J}'.$$

These two inequalities yield

$$\int_{B(c_i,\frac{1}{s})} e_\varepsilon(u_n) \geq \pi a(b_i)|\ln\varepsilon| + O(1) \quad \forall i \in \mathcal{J}'.$$

On the other hand, we have

$$F_a(u) \leq \pi\sum_{i=1}^{d} a(a_i)|\ln\varepsilon| + O(1) \leq \pi\sum_{i=1}^{d} a(b_i)|\ln\varepsilon| + O(1),$$

where we have used $\mathrm{dist}(a,b) \leq C\varepsilon^\gamma|\ln\varepsilon|$ (see Lemma 2.9). We finally get

$$\int_{B(c_i,\frac{1}{s})} |\nabla u_n|^2 \leq C.$$

Extracting a subsequence if necessary, there exists u_* such that

$$u_n \to u_* \text{ weakly in } H^1(\Omega \setminus \cup_{i=1}^d B(c_i,1/s)). \tag{5.6}$$

By standard diagonal extraction, we may find a subsequence such that this is true for any positive integer s. $|u_*| = 1$ follows from

$$\int_\Omega (1-|u_n|^2)^2 \leq C\varepsilon^2|\ln\varepsilon|^2.$$

Taking the cross product of (5.1) with u, we get

$$u_n \times [-\nabla\cdot(a(x)\nabla u_n) + 2ia(x)A_0\cdot\nabla u_n] = 0.$$

Using (5.5) and (5.6), we have from the above that u_* solves

$$u_* \times (\nabla\cdot(a(x)\nabla u_*)) = 0 \ \text{ in } \ \mathcal{D}'(\Omega\setminus\cup\{c_i\}).$$

Now it is easy to get the limit (1.12) from the above inequality. Theorem 1.2 is proved.

Acknowledgment. The authors would like to thank an anonymous referee for helpful suggestions.

REFERENCES

[1] A. Abrikosov, *On the magnetic properties of superconductivity of the second type*, Soviet Phys. JETP, 5 (1957), pp. 1174–1182.

[2] A. Aftalion, E. Sandier, and S. Serfaty, *Pinning phenomena in the Ginzburg-Landau model of superconductivity*, J. Math. Pures Appl. (9), 80 (2001), pp. 339–372.
[3] N. Andre, P. Bauman, and D. Phillips, *Mathematical Analysis of a Ginzburg-Landau System Related to Pinning*, preprint.
[4] N. Andre and I. Shafrir, *Asymptotic behavior of minimizers for the Ginzburg-Landau functional with weight*, Arch. Ration. Mech. Anal., 142 (1998), pp. 45–73 and 75–98.
[5] P. Bauman, D. Phillips, and Q. Tang, *Stable nucleation for the Ginzburg-Landau system with an applied magnetic field*, Arch. Ration. Mech. Anal., 142 (1998), pp. 1–43.
[6] A. Beaulieu and R. Hadiji, *On a class of Ginzburg-Landau equations with weight*, Panamer. Math. J., 5 (1995), pp. 1–33.
[7] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg-Landau Vortices*, Birkhäuser, Boston, 1994.
[8] F. Bethuel, H. Brezis, and F. Hélein, *Asymptotics for the minimization of a Ginzburg-Landau functional*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 123–148.
[9] F. Bethuel and T. Riviére, *Vortices for a variational problem related to superconductivity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 243–303.
[10] S. Chapman, Q. Du, and M. Gunzburger, *A model for variable thickness superconducting thin films*, Z. Angew. Math. Phys., 47 (1996), pp. 410–431,
[11] S. Chapman, Q. Du, and M. Gunzburger, *A Ginzburg-Landau type model of superconducting/normal junctions including Josephson junctions*, European J. Appl. Math., 6 (1995), pp. 97–114.
[12] S. Chapman and G. Richardson, *Vortex pinning by inhomogeneities in type*-II *superconductors*, Phys. D, 108 (1997), pp. 397–407.
[13] S. Ding, *Renormalized energy with vortices pinning effects*, J. Partial Differential Equations, 13 (2000), pp. 341–360.
[14] S. Ding and Z. Liu, *Pinning of vortices for a variational problem related to the superconducting thin films having variable thickness*, J. Partial Differential Equations, 10 (1997), pp. 174–192.
[15] S. Ding, Z. Liu, and W. Yu, *Pinning of vortices for the Ginzburg-Landau functional with variable coefficient*, Appl. Math. J. Chinese Univ. Ser. B, 12 (1997), pp. 77–88.
[16] Q. Du, M. Gunzburger, and J. Peterson, *Analysis and approximation of the Ginzburg-Landau model of superconductivity*, SIAM Rev., 34 (1992), pp. 54–81.
[17] G. Lasher, *Mixed states of type*-I *superconducting films in a perpendicular magnetic field*, Phys. Rev. (2), 154 (1967), pp. 345–348.
[18] F.-H. Lin and Q. Du, *Ginzburg–Landau vortices: Dynamics, pinning and hysteresis*, SIAM J. Math. Anal., 28 (1997), pp. 1265–1293.
[19] J. Livingston and W. Desoro, *The Intermediate State in Type*-I *Superconductors*, in Superconductivity, R. Parks, ed., Marcel Dekker, New York, 1969, pp. 1235–1281.
[20] K. Maki, *Fluxoid structure in superconducting films*, Ann. Physics, 34 (1965), pp. 363–376.
[21] E. Sandier and S. Serfaty, *Global minimizers for the Ginzburg-Landau functional below the first critical magnetic field*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 119–145.
[22] E. Sandier and S. Serfaty, *On the energy of type*-II *superconductors in the mixed phase*, Rev. Math. Phys., 12 (2000), pp. 1219–1257.
[23] S. Serfaty, *Local minimizer for the Ginzburg-Landau energy near critical magnetic field: Part* I, Commun. Contemp. Math., 1 (1999), pp. 213–254.
[24] S. Serfaty, *Local minimizer for the Ginzburg-Landau energy near critical magnetic field: Part* II, Commun. Contemp. Math., 1 (1999), pp. 295–333.
[25] S. Serfaty, *Stable configurations in superconductivity: Uniqueness, multiplicity, and vortex-nucleation*, Arch. Ration. Mech. Anal., 149 (1999), pp. 329–365.
[26] M. Tinkham, *Introduction to Superconductivity*, 2nd ed., McGraw-Hill, New York, 1994.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 257–272

THE LIFSHITZ–SLYOZOV–WAGNER EQUATION WITH CONSERVED TOTAL VOLUME*

PHILIPPE LAURENÇOT†

Abstract. The Lifshitz–Slyozov–Wagner theory of coarsening (Ostwald ripening) in alloys describes the time evolution of the sizes of the grains of a new phase growing by diffusional mass transfer from a supersaturated solid solution. The volume distribution function of the grains obeys a nonlinear transport equation with a nonlocal nonlinearity. Global existence of solutions is obtained for a large class of data including the ones derived by Lifshitz and Slyozov [*J. Phys. Chem. Solids*, 19 (1961), pp. 35–50] and Wagner [*Z. Elektrochem.*, 65 (1961), pp. 581–591], and uniqueness of these solutions is proved in some cases.

Key words. Lifshitz–Slyozov model, Ostwald ripening, existence, uniqueness

AMS subject classifications. 35L60, 82C21

PII. S0036141001387471

1. Introduction. The theory of coarsening (Ostwald ripening) in alloys describes the late stages of the formation and growth of grains of a new phase from a supersaturated solid solution. During these stages, no new grains can form and the determining process is the growth of the grains by diffusional mass exchange [6, 11]. More precisely, the grains of the new phase larger than some critical size grow at the expense of smaller ones, the critical size varying in time as a function of the degree of supersaturation. A mean-field approach for this process has been formulated by Lifshitz and Slyozov [6] and Wagner [11]. The resulting model consists of an evolution equation for the volume distribution function f of the grains coupled with the equation of the conservation of matter and reads, for spherical grains,

$$f_t + (\mathcal{V}\ f)_x = 0, \quad (t,x) \in (0,+\infty) \times (0,+\infty),$$
$$u(t) + A \int_0^\infty x\ f(t,x)\ dx = Q, \quad t \in (0,+\infty).$$

Here $x \in (0,+\infty)$ is the volume of the grains, $t \in (0,+\infty)$ is the time variable, Q is the total initial supersaturation, and A is a physical constant [6]. Finally $\mathcal{V} = \mathcal{V}(t,x)$ denotes the rate of growth of the grains and is determined by the mechanism of mass transfer between the grains, e.g., volume diffusion [6, 11] or grain-boundary diffusion [9]. In general one has $\mathcal{V}(t,x) = k(x)u(t) - q(x)$, where k and q are computed from the modeling of the mechanism of mass transfer between the grains [6, 9, 11], and the Lifshitz–Slyozov–Wagner equation reads

$$f_t + ((ku - q)\ f)_x = 0, \quad (t,x) \in (0,+\infty) \times (0,+\infty), \tag{1}$$
$$u(t) + A \int_0^\infty x\ f(t,x)\ dx = Q, \quad t \in (0,+\infty). \tag{2}$$

For instance, in the model considered in [6], the grains are assumed to be widely separated spheres evolving in a quasi-static diffusion field. The interaction between

*Received by the editors April 6, 2001; accepted for publication (in revised form) April 11, 2002; published electronically October 8, 2002.
http://www.siam.org/journals/sima/34-2/38747.html
†Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, Université Paul Sabatier—Toulouse 3, 118 route de Narbonne, F-31062 Toulouse cedex 4, France (laurenco@mip.ups-tlse.fr).

the grains is ignored and the diffusion field is taken to be close to the degree of supersaturation far from the grains. Assuming the growth rate of the grains to be proportional to the mass flux at the grain boundary, it can be computed explicitly [6] and reads, in dimensionless form,

$$\mathcal{V}(t,x) = 3\ \left(x^{1/3}\ u(t) - 1 \right),$$

that is, $k(x) = 3\ x^{1/3}$ and $q(x) = 3$, $x \in (0,+\infty)$.

For very dilute solutions a physically relevant assumption at large times is that the variation of the degree of supersaturation is small during the time evolution and the equation of conservation of matter (2) becomes [11]

$$\int_0^\infty x\ f(t,x)\ dx = \text{ const.}, \quad t \in (0,+\infty).$$

In that case the function u is determined by requiring that the solution f to (1) comply with the above conservation law, that is,

$$u(t)\ \int_0^\infty k(x)\ f(t,x)\ dx = \int_0^\infty q(x)\ f(t,x)\ dx, \quad t \in (0,+\infty).$$

The aim of this work is thus to investigate the existence and uniqueness of weak solutions to the initial value problem

$$f_t + ((ku - q)\ \ f)_x = 0, \quad (t,x) \in (0,+\infty) \times (0,+\infty), \tag{3}$$

$$u(t)\ \int_0^\infty k(x)\ f(t,x)\ dx = \int_0^\infty q(x)\ f(t,x)\ dx, \quad t \in (0,+\infty), \tag{4}$$

$$f(0,x) = f_0(x), \quad x \in (0,+\infty). \tag{5}$$

The initial value problem (3)–(5) is a nonlinear transport equation with a nonlocal nonlinearity. Observe, however, that the main difference between (2) and (4) is that u is a linear functional of f in the former, while it is a nonlinear functional of f in the latter. It is thus expected that the initial value problem (3)–(5) will be more delicate to handle than the initial value problem (3), (2), and (5). Still, existence and uniqueness of measure-valued solutions are proved by Niethammer and Pego when k and q are given by

$$k(x) = 3\ x^{1/3} \quad \text{and} \quad q(x) = 3, \quad x \in (0,+\infty), \tag{6}$$

the initial datum f_0 being a probability measure with compact support [7]. The extension of this result to an arbitrary probability measure is performed in [8]. To our knowledge these are the only available existence and uniqueness results for (3)–(5), and the purpose of this work is to investigate the existence and uniqueness of solutions to (3)–(5) for a larger class of functions k and q, including the ones derived in [6], which are given by (6), and the ones derived in [11], namely,

$$k(x) = \frac{a\ x^{2/3}}{c\ x^{1/3} + d} \quad \text{and} \quad q(x) = \frac{b\ x^{1/3}}{c\ x^{1/3} + d}, \quad x \in (0,+\infty), \tag{7}$$

where a, b, c, and d are positive real numbers. Our analysis relies on different arguments than the ones developed in [7] and actually includes the following typical example (which generalizes the Lifshitz–Slyozov case (6)):

$$k(x) = a\ x^{\alpha} \quad \text{and} \quad q(x) = b\ x^{\beta}, \quad x \in (0, +\infty), \tag{8}$$

where $0 \le \beta < \alpha \le 1$ and a, b are positive real numbers. We will, however, restrict ourselves to nonnegative and integrable initial data with finite first moment and do not consider the case of measures. Before going further, let us mention that the initial value problem (3), (2), and (5) has been studied recently, and existence and uniqueness of measure-valued and integrable solutions have been obtained in [1, 7] and [1, 4], respectively.

We now briefly outline the contents of the paper and sketch the main ideas of the existence proof as well. In the next section we state the assumptions on the data k, q, and f_0 together with our main results. The case where the functions k and q are given by (8) is included in our analysis, and we prove the existence of a solution to (3)–(5) when $0 \le \beta < \alpha \le 1$ for any nonnegative and integrable initial datum f_0 with finite first moment. In addition, this solution is shown to be unique if $\beta > 0$ or $(\alpha, \beta) = (1, 0)$. Thus, our uniqueness result unfortunately does not apply to the Lifshitz–Slyozov case (6). Nevertheless our results apply when k and q are given by (7) and provide also the existence and uniqueness of a solution to (3)–(5) in that case. The existence proof is inspired by the derivation of (3)–(5) performed in [11] and may be seen somehow as a penalization method. More precisely, we consider $\varepsilon \in (0, 1)$ and denote by $(f^{\varepsilon}, u^{\varepsilon})$ the solution to (3), (2), and (5), where we have chosen

$$A_{\varepsilon} = \varepsilon^{-1} \quad \text{and} \quad Q_{\varepsilon} = \varepsilon^{-1} \int_0^{\infty} x\ f_0(x)\ dx.$$

Then (2) reads

$$\varepsilon\ u^{\varepsilon}(t) = \int_0^{\infty} x\ f_0(x)\ dx - \int_0^{\infty} x\ f^{\varepsilon}(t,x)\ dx, \quad t \in [0, +\infty),$$

and it is easily seen (at least formally) that the above equation yields (4) as $\varepsilon \to 0$. We might thus expect that the sequence $(f^{\varepsilon}, u^{\varepsilon})$ will converge to a solution to (3)–(5), and this turns out to be true as we shall see below. We thus recall in section 3.1 some results for (3), (2), and (5) previously obtained in [4], namely, the existence of solutions together with some estimates which will be needed later. Section 3.2 is devoted to the main step of the existence proof, namely, an L^{∞}-bound for u^{ε} which is uniform with respect to $\varepsilon \in (0, 1)$. Thanks to this bound, we may argue as in [4] and prove that (f^{ε}) enjoys some weak compactness properties in $L^1(0, +\infty; x dx)$ with the help of a refined version of the de la Vallée–Poussin theorem [5]. Equicontinuity with respect to time then follows from (3) and allows us to complete the proof of the existence result in section 3.3. Uniqueness of solutions to (3)–(5) is investigated in the final section and requires stronger assumptions on the functions k and q.

Remark. The above choice of Q_{ε} entails that $u^{\varepsilon}(0) = 0$ and is made throughout the paper for simplicity. However, the convergence of $(f^{\varepsilon}, u^{\varepsilon})$ shown in section 3 is still valid with $Q_{\varepsilon} = Q_{\varepsilon}^0/\varepsilon$, (Q_{ε}^0) being a nondecreasing sequence satisfying

$$\lim_{\varepsilon \to 0} Q_{\varepsilon}^0 = \int_0^{\infty} x\ f_0(x)\ dx.$$

2. Main results. We first describe the class of functions f_0, k, and q to be considered in this paper. More precisely, we assume that the data f_0, k, and q enjoy the following properties:

$$f_0 \in L^1(\mathbb{R}_+;(1+x)dx) \quad \text{and} \quad f_0 \ge 0 \ \text{ a.e. in } \ \mathbb{R}_+, \tag{9}$$

where $\mathbb{R}_+ = (0,+\infty)$.

$$\begin{cases} \text{The function } k \text{ is a nonnegative function in } \mathcal{C}([0,+\infty)) \cap \mathcal{C}^1(\mathbb{R}_+) \\ \text{satisfying } k(0)=0,\ k(r)>0 \text{ if } r>0 \text{ and} \\ \qquad k' \in L^\infty(1,+\infty) \quad \text{and} \quad k' \ge 0, \\ \qquad r \mapsto \dfrac{k(r)}{r} \ \text{ is nonincreasing on } \ \mathbb{R}_+. \end{cases} \tag{10}$$

$$\begin{cases} \text{The function } q \text{ is a nonnegative function in } \mathcal{C}([0,+\infty)) \cap \mathcal{C}^1(\mathbb{R}_+) \\ \text{and satisfies} \\ \qquad q' \in L^\infty(1,+\infty) \quad \text{and} \quad q' \ge 0. \end{cases} \tag{11}$$

In other words the functions k and q are Lipschitz continuous functions for large values of x and might be less regular near $x=0$ but are nondecreasing. Note also that since $k(0)=0$ and q is nonnegative, no boundary condition is needed at $x=0$ to solve (3).

We also assume that for every $U \ge 0$, there exists $x_U \in (0,1]$ such that

$$U\ k(x) - q(x) \le -x\ q'(x), \quad x \in (0,x_U]. \tag{12}$$

In particular, q' being nonnegative by (11), we infer from (12) that

$$\lim_{x\to 0} \frac{q(x)}{k(x)} = +\infty. \tag{13}$$

Let us point out here that the functions k and q given by (6) (see [6]) and (7) (see [11]) fulfill the assumptions (10)–(12) and the functions k and q given by (8) as well (since $0 \le \beta < \alpha \le 1$).

Remark. The assumption that k and q are nondecreasing may actually be relaxed, and the results presented below are also true for Lipschitz continuous perturbations of nondecreasing functions k and q. We restrict ourselves, however, to the framework described above for simplicity and refer to [4], where this more general class of data is considered for the initial value problem (3), (2), and (5).

We are now in a position to state our existence result.

THEOREM 1. *Consider a function f_0 satisfying* (9) *and assume that the functions k and q enjoy the properties* (10)–(12). *There are at least a couple of nonnegative functions (f,u) satisfying*

$$\begin{cases} f \in \mathcal{C}([0,t];L^1(\mathbb{R}_+;xdx)) \cap L^\infty(0,t;L^1(\mathbb{R}_+)), \\ u \in L^\infty(0,t) \end{cases} \tag{14}$$

and

$$\int_0^\infty f(t,x)\ g(x)\ dx = \int_0^\infty f_0(x)\ g(x)\ dx + \int_0^t \int_0^\infty g_x(x)\ \mathcal{V}(s,x)\ f(s,x)\ dxds \tag{15}$$

for each $t \in \mathbb{R}_+$ *and* $g \in \mathcal{C}_0^\infty(\mathbb{R}_+)$, *where*

$$\mathcal{V}(t,x) = k(x)\ u(t) - q(x), \quad x \in \mathbb{R}_+, \tag{16}$$

$$u(t) \int_0^\infty k(x)\ f(t,x)\ dx = \int_0^\infty q(x)\ f(t,x)\ dx, \tag{17}$$

or, equivalently,

$$\int_0^\infty x\ f(t,x)\ dx = \int_0^\infty x\ f_0(x)\ dx. \tag{18}$$

Note that (15) makes sense since the continuity of k, q, and (14) ensure that $\mathcal{V} \in L^\infty((0,T) \times \mathcal{K})$ for every compact subset $\mathcal{K}$ of $\mathbb{R}_+$ and $T \in \mathbb{R}_+$. Let us also mention here that, in the proof of Theorem 1, we first obtain that $f \in \mathcal{C}([0,+\infty); w - L^1(\mathbb{R}_+; xdx))$. Here we use the following notation: if X is a Banach space and $T \in (0,+\infty]$, then $\mathcal{C}([0,T); w - X)$ denotes the space of weakly continuous functions from $[0,T)$ in X. The time continuity (14) of f in the strong topology of $L^1(\mathbb{R}_+; xdx)$ then follows from (3) by arguments similar to those of [3, sections II.1 and II.2].

If we strengthen the assumptions on the data k and q, we are able to show that there is only one solution to (3)–(5) with the properties stated in Theorem 1.

THEOREM 2. *Assume that* f_0, k, *and* q *fulfill* (9)–(12) *and that*

$$\sup_{x \in (0,+\infty)} (U\ k'(x) - q'(x)) < +\infty \tag{19}$$

for each $U \in \mathbb{R}_+$. *Then there are a unique couple of nonnegative functions* (f,u) *satisfying* (14)–(17).

Clearly the functions k and q given by (8) satisfy (19) only if $0 < \beta < \alpha \leq 1$ or $(\alpha,\beta) = (1,0)$, which unfortunately excludes the Lifshitz–Slyozov case (6). The assumption (19) is also fulfilled in the Wagner case (7) and guarantees that $\mathcal{V}_x$ is bounded from above on $(0,T) \times \mathbb{R}_+$ for each $T > 0$.

From now on we assume that f_0, k, and q are given functions satisfying (9)–(12). Since $f \equiv 0$ is clearly a solution to (3)–(5) with initial datum $f_0 \equiv 0$, we further assume that $f_0 \not\equiv 0$ and put

$$M_0 := \int_0^\infty x\ f_0(x)\ dx > 0.$$

In the following we denote by C any positive constant depending only on f_0, k, and q. The dependence of C upon additional parameters will be indicated explicitly.

3. Existence.

3.1. The approximating equation. For $\varepsilon \in (0,1)$ we put

$$A_\varepsilon = \varepsilon^{-1} \quad \text{and} \quad Q_\varepsilon = \varepsilon^{-1} \int_0^\infty x\ f_0(x)\ dx. \tag{20}$$

Owing to (9)–(12) and (20), we are in a position to apply [4, Theorem 2.2, Propositions 3.1 and 3.3] to obtain the existence of a weak solution to (3), (2), and (5) with initial datum f_0 and $(A_\varepsilon, Q_\varepsilon)$ instead of (A,Q). More precisely, we have the following result.

PROPOSITION 3. *For $\varepsilon \in (0,1)$ there is a nonnegative function*

$$f^\varepsilon \in \mathcal{C}([0,+\infty); L^1(\mathbb{R}_+; x dx)) \cap L^\infty(0,+\infty; L^1(\mathbb{R}_+)) \tag{21}$$

satisfying for each $t \in \mathbb{R}_+$ and $g \in \mathcal{C}_0^\infty(\mathbb{R}_+)$

$$A_\varepsilon \int_0^\infty x\ f^\varepsilon(t,x)\ dx \le Q_\varepsilon, \tag{22}$$

$$\begin{aligned} \int_0^\infty f^\varepsilon(t,x)\ g(x)\ dx &= \int_0^\infty f_0(x)\ g(x)\ dx \\ &\quad + \int_0^t \int_0^\infty g_x(x)\ \mathcal{V}^\varepsilon(s,x)\ f^\varepsilon(s,x)\ dxds, \end{aligned} \tag{23}$$

where

$$\mathcal{V}^\varepsilon(t,x) = k(x)\ u^\varepsilon(t) - q(x), \quad x \in \mathbb{R}_+, \tag{24}$$

$$u^\varepsilon(t) + A_\varepsilon \int_0^\infty x\ f^\varepsilon(t,x)\ dx = Q_\varepsilon. \tag{25}$$

In addition,

$$\int_0^\infty f^\varepsilon(t,x)\ dx \le \int_0^\infty f^\varepsilon(s,x)\ dx \le \int_0^\infty f_0(x)\ dx, \quad 0 \le s \le t. \tag{26}$$

Notice that (22) warrants that u^ε is nonnegative, while (25) also reads

$$\varepsilon\ u^\varepsilon(t) + \int_0^\infty x\ f^\varepsilon(t,x)\ dx = \int_0^\infty x\ f_0(x)\ dx, \quad t \in [0,+\infty). \tag{27}$$

From the analysis of [4] we also deduce some estimates on the moments of f^ε, together with some integrability properties. We first introduce the set $\mathcal{J}_\infty$ of nonnegative and convex functions $j : [0,+\infty) \longrightarrow [0,+\infty)$ such that

$$\begin{cases} j \in \mathcal{C}^1([0,+\infty)) \cap W^{2,\infty}_{\rm loc}(\mathbb{R}_+) \text{ with } j(0) = 0 \text{ and } j'(0) \ge 0,\ j' \text{ is a} \\ \text{concave function on } [0,+\infty) \text{ and} \\ \qquad \displaystyle\lim_{r\to+\infty} j'(r) = \lim_{r\to+\infty} \frac{j(r)}{r} = +\infty. \end{cases} \tag{28}$$

The next result follows at once from [4, Lemma 3.4] and provides a control on the propagation of moments of f^ε.

LEMMA 4. *Let $\varepsilon \in (0,1)$ and $T \in \mathbb{R}_+$. Assume that there are a function $j \in \mathcal{J}_\infty$ and a real number U such that*

$$M = \int_0^\infty j(x)\ f_0(x)\ dx < +\infty \quad \text{and} \quad \sup_{t\in[0,T]} \{u^\varepsilon(t)\} \le U. \tag{29}$$

There is a constant K_1 depending only on k, q, f_0, j, M, U, x_U, and T such that

$$\int_0^\infty j(x)\ f^\varepsilon(t,x)\ dx \le K_1, \quad t \in [0,T]. \tag{30}$$

Notice that since j is superlinear at infinity, Lemma 4 allows us to control the behavior of f^ε for large values of x. As for the local behavior of f^ε we have the following result, which is a consequence of [4, Lemma 3.5].

LEMMA 5. *Let $\varepsilon \in (0,1)$ and $T \in \mathbb{R}_+$. Assume that there are a function $j \in \mathcal{J}_\infty$ and a real number U such that*

$$M = \int_0^\infty j\left(f_0(x)\right)\ x\ dx < +\infty \quad \text{and} \quad \sup_{t\in[0,T]} \{u^\varepsilon(t)\} \le U. \tag{31}$$

There is a constant K_2 depending only on k, q, f_0, j, M, U, x_U, and T such that

$$\int_0^\infty j\left(f^\varepsilon(t,x)\right)\ \min(x,1)\ dx \le K_2, \quad t \in [0,T]. \tag{32}$$

Here again the superlinearity of j at infinity ensures that $f^\varepsilon(t)$ cannot concentrate on a small measurable subset of $\mathbb{R}_+$ and thus excludes the formation of Dirac masses. We thus conclude from the previous results that if f_0 enjoys the integrability properties (29) and (31) and if the sequence (u^ε) is uniformly bounded in $L^\infty(0,T)$, the sequence $(f^\varepsilon(t))$ is uniformly integrable in $L^1(\mathbb{R}_+; xdx)$ for $t \in [0,T]$, whence it is weakly compact in $L^1(\mathbb{R}_+; xdx)$ by the Dunford–Pettis theorem. Therefore an L^∞-bound on u^ε seems to be an important step towards the proof of Theorem 1 and is derived in the next section.

3.2. An L^∞-estimate for u^ε. We now turn to the cornerstone of the proof of Theorem 1 and prove the following result.

LEMMA 6. *Let $T \in \mathbb{R}_+$. There are $\varepsilon(T) \in (0,1)$ and $C(T)$ such that there holds*

$$u^\varepsilon(t) \le C(T)$$

for every $\varepsilon \in (0, \varepsilon(T))$ and $t \in [0,T]$.

Proof. The proof of Lemma 6 actually splits into two parts and depends on the compactness or noncompactness of the support of f_0.

Case 1. We first consider the case where

$$\int_x^\infty f_0(y)\ dy > 0 \tag{33}$$

for every $x \in \mathbb{R}_+$ (i.e., f_0 is not compactly supported). It follows from (23), (24), and (25) that

$$\frac{du^\varepsilon}{dt}(t) + \frac{1}{\varepsilon}\ \left(\int_0^\infty k(x)\ f^\varepsilon(t,x)\ dx\right)\ u^\varepsilon(t) = \frac{1}{\varepsilon}\ \int_0^\infty q(x)\ f^\varepsilon(t,x)\ dx.$$

Owing to (11), there holds $q(x) \le C\ (1+x)$ for $x \in \mathbb{R}_+$, and the right-hand side of the above equality can be bounded from above with the help of (25) and (26) by

$$\frac{1}{\varepsilon}\ \int_0^\infty q(x)\ f^\varepsilon(t,x)\ dx \le \frac{C}{\varepsilon}\ \int_0^\infty (1+x)\ f^\varepsilon(t,x)\ dx \le \frac{C}{\varepsilon}\ \int_0^\infty (1+x)\ f_0(x)\ dx.$$

Therefore,

$$\frac{du^\varepsilon}{dt}(t) + \frac{1}{\varepsilon}\ \left(\int_0^\infty k(x)\ f^\varepsilon(t,x)\ dx \right)\ u^\varepsilon(t) \le \frac{C}{\varepsilon}. \tag{34}$$

We now introduce the function F^ε defined by

$$F^\varepsilon(t,x) = \int_x^\infty f^\varepsilon(t,y)\ dy, \quad (t,x) \in [0,+\infty) \times \mathbb{R}_+.$$

Owing to (21) and (23), we have $F_t^\varepsilon = \mathcal{V}^\varepsilon\ f^\varepsilon$, and the nonnegativity of k, u^ε, and f^ε further entails that $F_t^\varepsilon \ge -q\ f^\varepsilon$. Since $q(x) \le C\ (1+x)$ for $x \in \mathbb{R}_+$ by (11) and $f^\varepsilon = -F_x^\varepsilon$, we end up with

$$F_t^\varepsilon \ge C\ (1+x)\ F_x^\varepsilon,$$

whence

$$F^\varepsilon(t,x) \ge F^\varepsilon\left(0, (1+x)\ e^{Ct} - 1\right), \quad (t,x) \in [0,+\infty) \times \mathbb{R}_+.$$

Since k satisfies (10) with $k(0) = 0$, we deduce from the above estimate that

$$\begin{aligned} \int_0^\infty k(x)\ f^\varepsilon(t,x)\ dx &\ge \int_{e^{-Ct}}^\infty k(x)\ f^\varepsilon(t,x)\ dx \\ &\ge k\left(e^{-Ct}\right)\ F^\varepsilon\left(t, e^{-Ct}\right) \\ &\ge k\left(e^{-Ct}\right)\ F^\varepsilon\left(0, e^{Ct}\right) \\ &\ge k\left(e^{-CT}\right)\ \int_{e^{CT}}^\infty f_0(y)\ dy. \end{aligned}$$

Recalling (34), we finally obtain

$$\frac{du^\varepsilon}{dt}(t) + \frac{\delta(T)}{\varepsilon}\ u^\varepsilon(t) \le \frac{C}{\varepsilon}$$

with

$$\delta(T) := k\left(e^{-CT}\right)\ \int_{e^{CT}}^\infty f_0(y)\ dy.$$

Thanks to (10) and (33) we have $\delta(T) > 0$, and the differential inequality satisfied by u^ε yields

$$u^\varepsilon(t) \le \frac{C}{\delta(T)}, \quad t \in [0,T].$$

Recall that $u^\varepsilon(0) = 0$ by (20) and (25). We have thus proved Lemma 6 for noncompactly supported initial data (with $\varepsilon(T) = 1$).

Case 2. We now turn to the case of a compactly supported initial datum f_0 and assume that $f_0(x) = 0$ a.e. in $(R_0, +\infty)$ for some $R_0 > 0$. We first notice that the previous proof does not work in that case as $\delta(T)$ vanishes for T large enough. Since (3) is a transport equation, we actually expect $f^\varepsilon(t)$ to be compactly supported for each $t \geq 0$, and the proof of Lemma 6 relies on an estimate of the growth of the support of f^ε, which we derive now.

LEMMA 7. *There is a unique couple* $(R_\varepsilon, \tau_\varepsilon)$ *in* $\mathcal{C}([0,+\infty)) \times (0,+\infty]$ *satisfying* $R_\varepsilon(0) = R_0$, $R_\varepsilon(t) > 0$ *if* $t \in [0, \tau_\varepsilon)$, $R_\varepsilon \in \mathcal{C}^1([0,\tau_\varepsilon))$, *and*

$$\begin{cases} \dfrac{dR_\varepsilon}{dt}(t) = \mathcal{V}^\varepsilon\left(t, R_\varepsilon(t)\right) & \text{if} \quad t \in [0, \tau_\varepsilon), \\ \\ R_\varepsilon(t) = 0 & \text{if} \quad t \geq \tau_\varepsilon. \end{cases} \tag{35}$$

In addition, the support of $f^\varepsilon(t)$ *is included in* $[0, R_\varepsilon(t)]$ *for each* $t \geq 0$.

Proof of Lemma 7. First we remark that $\mathcal{V}^\varepsilon$ is continuous on $[0,+\infty)^2$ and Lipschitz continuous with respect to x on compact subsets of $[0,+\infty)\times\mathbb{R}_+$. Since $R_0 > 0$, classical results ensure that there is a unique maximal solution $R_\varepsilon \in \mathcal{C}^1([0,\tau_\varepsilon);\mathbb{R}_+)$ to

$$\frac{dR_\varepsilon}{dt}(t) = \mathcal{V}^\varepsilon\left(t, R_\varepsilon(t)\right), \quad R_\varepsilon(0) = R_0,$$

and we have the following alternative: either $\tau_\varepsilon = +\infty$, or $\tau_\varepsilon < +\infty$ and the only possible cluster points of $R_\varepsilon(t)$ as $t \to \tau_\varepsilon-$ are 0 and $+\infty$. In the latter case, notice that (10), (11), and (25) entail that

$$\frac{dR_\varepsilon}{dt}(t) \leq C\ Q_\varepsilon\ \left(1 + R_\varepsilon(t)\right), \quad t \in [0, \tau_\varepsilon),$$

which excludes the possibility of blow-up as $t \to \tau_\varepsilon-$. Consequently $R_\varepsilon(t)$ converges to 0 as $t \to \tau_\varepsilon-$, and we extend R_ε to $[0,+\infty)$ by putting $R_\varepsilon(t) = 0$ for $t \geq \tau_\varepsilon$. The estimate for the support of $f^\varepsilon(t)$ then follows by standard arguments if $t \in [0,\tau_\varepsilon)$. If $\tau_\varepsilon < +\infty$, we further obtain that $f^\varepsilon(\tau_\varepsilon) \equiv 0$, whence $f^\varepsilon(t) \equiv 0$ for $t \geq \tau_\varepsilon$ by (26), and the proof of Lemma 7 is complete. □

We are now in a position to complete the proof of Lemma 6. Let $T \in \mathbb{R}_+$ and consider $t \in [0,T]$ such that $t < \tau_\varepsilon$. We infer from (10) and (27) that

$$\begin{aligned} \int_0^\infty k(x)\ f^\varepsilon(t,x)\ dx &= \int_0^{R_\varepsilon(t)} \frac{k(x)}{x}\ x\ f^\varepsilon(t,x)\ dx \\ &\geq \frac{k\left(R_\varepsilon(t)\right)}{R_\varepsilon(t)}\ \int_0^{R_\varepsilon(t)} x\ f^\varepsilon(t,x)\ dx \\ &\geq \frac{k\left(R_\varepsilon(t)\right)}{R_\varepsilon(t)}\ \int_0^{\infty} x\ f^\varepsilon(t,x)\ dx \\ &\geq \frac{k\left(R_\varepsilon(t)\right)}{R_\varepsilon(t)}\ \left(M_0 - \varepsilon\ u^\varepsilon(t)\right). \end{aligned}$$

Recalling (34) and (35), we obtain the following system of differential inequalities for $(u^\varepsilon, R_\varepsilon)$:

$$\frac{du^\varepsilon}{dt}(t) + \frac{k\left(R_\varepsilon(t)\right)}{R_\varepsilon(t)}\ \left(\frac{M_0}{\varepsilon} - u^\varepsilon(t)\right)\ u^\varepsilon(t) \leq \frac{C}{\varepsilon}, \tag{36}$$

$$\frac{dR_\varepsilon}{dt}(t) \leq \frac{k\left(R_\varepsilon(t)\right)}{R_\varepsilon(t)}\ u^\varepsilon(t)\ R_\varepsilon(t), \tag{37}$$

which is valid for $t \in [0,T] \cap [0,\tau_\varepsilon)$. Since $u^\varepsilon(0) = 0$ by (20), we have

$$\sigma_\varepsilon := \sup \left\{ t \in [0,T] \cap [0,\tau_\varepsilon),\ \ u^\varepsilon(s) \le \frac{M_0}{2\ \varepsilon} \ \text{ for } \ s \in [0,t) \right\} > 0.$$

Assume for contradiction that $\sigma_\varepsilon < \min\{T, \tau_\varepsilon\}$. On the one hand, we infer from (36) that

$$\frac{du^\varepsilon}{dt}(t) + \frac{M_0}{2\ \varepsilon}\ \frac{k\left(R_\varepsilon(t)\right)}{R_\varepsilon(t)}\ u^\varepsilon(t) \le \frac{C}{\varepsilon} \tag{38}$$

for $t \in [0,\sigma_\varepsilon]$, whence, after integration,

$$\begin{aligned} \frac{M_0}{2\ \varepsilon}\ \int_0^t \frac{k\left(R_\varepsilon(s)\right)}{R_\varepsilon(s)}\ u^\varepsilon(s)\ ds &\le \frac{C\ t}{\varepsilon}, \\ \int_0^t \frac{k\left(R_\varepsilon(s)\right)}{R_\varepsilon(s)}\ u^\varepsilon(s)\ ds &\le C\ t, \end{aligned}$$

since u^ε is nonnegative and $u^\varepsilon(0) = 0$. On the other hand, the positivity of R_ε on $[0,\sigma_\varepsilon]$ and (37) entails that

$$R_\varepsilon(t) \le R_0\ \exp\left\{ \int_0^t \frac{k\left(R_\varepsilon(s)\right)}{R_\varepsilon(s)}\ u^\varepsilon(s)\ ds \right\}, \quad t \in [0,\sigma_\varepsilon].$$

Combining the above two estimates finally yields

$$R_\varepsilon(t) \le R_0\ e^{Ct} \le C(T), \quad t \in [0,\sigma_\varepsilon].$$

Recalling that $r \mapsto k(r)/r$ is nonincreasing by (10), we may use the above upper bound on R_ε to estimate the second term of the left-hand side of (38) from below and obtain

$$\frac{du^\varepsilon}{dt}(t) + \frac{C(T)}{\varepsilon}\ \ u^\varepsilon(t) \le \frac{C}{\varepsilon}, \quad t \in [0,\sigma_\varepsilon].$$

The Gronwall lemma then ensures that

$$u^\varepsilon(t) \le C(T), \quad t \in [0,\sigma_\varepsilon],$$

and a contradiction for ε small enough. Therefore $\sigma_\varepsilon = \min\{T,\tau_\varepsilon\}$ for ε small enough, and the above computation entails that

$$u^\varepsilon(t) \le C(T), \quad t \in [0, \min\{T,\tau_\varepsilon\}]. \tag{39}$$

We next argue again by contradiction to show that $\tau_\varepsilon > T$ for ε small enough. Otherwise $\tau_\varepsilon \le T$ for $\varepsilon \in (0,1)$ and $f^\varepsilon(\tau_\varepsilon) \equiv 0$. Therefore $u^\varepsilon(\tau_\varepsilon) = M_0/\varepsilon$ by (25), while (39) ensures that $u^\varepsilon(\tau_\varepsilon) \le C(T)$, whence we obtain a contradiction for ε small enough. Consequently $\tau_\varepsilon > T$ for ε small enough, and Lemma 6 then follows from (39). □

3.3. Proof of Theorem 1. We are now in a position to complete the proof of Theorem 1. As already mentioned, we proceed along the lines of the proof of [4, Theorem 2.2] and aim to prove that the sequence (f^ε) is relatively weakly compact in $L^1((0,T)\times\mathbb{R}_+; x dx dt)$ for each $T\in\mathbb{R}_+$. For that purpose we need to be able to control the behavior of the sequence (f^ε) for large values of x and on small measurable subsets of $(0,T)\times\mathbb{R}_+$. Concerning the latter it is equivalent to obtain an upper bound on the L^1-norm of $j(f^\varepsilon)$ for some function j which is superlinear for large values of its argument. Thanks to the L^∞-bound on u^ε obtained in Lemma 6, both results will follow from Lemma 4 and Lemma 5, respectively. We first recall a refined version of the de la Vallée–Poussin theorem [5, Proposition I.1.1].

THEOREM 8. *If $(\Omega,\mathcal{B},\mu)$ is a measured space and $w\in L^1(\Omega,\mathcal{B},\mu)$, there exists a function $j\in\mathcal{J}_\infty$ (depending only on w) such that*

$$j\left(|w|\right)\in L^1(\Omega,\mathcal{B},\mu).$$

Remark. Theorem 8 is a classical result when $\mu(\Omega)<\infty$ (see, e.g., [2, p. 38]), except for the possibility of choosing j' concave. This last fact has been noticed in [5].

Owing to (9), we may apply Theorem 8 to conclude that there are two functions j_1 and j_2 in $\mathcal{J}_\infty$ such that

$$M:=\int_0^\infty j_1(x)\ f_0(x)\ dx+\int_0^\infty j_2\left(f_0(x)\right)\ x\ dx<\infty. \tag{40}$$

We fix $T\in\mathbb{R}_+$. Owing to (40) and Lemma 6, we may apply Lemmas 4 and 5 and conclude that there holds

$$\int_0^\infty j_1(x)\ f^\varepsilon(t,x)\ dx+\int_0^\infty j_2\left(f^\varepsilon(t,x)\right)\ \min\left(x,1\right)\ dx\le C(T)$$

for every $\varepsilon\in(0,\varepsilon(T))$ and $t\in(0,T)$. Since both j_1 and j_2 are superlinear at infinity, we infer from the above estimates and the Dunford–Pettis theorem that there is a weakly compact subset $\mathcal{K}(T)$ of $L^1(\mathbb{R}_+;xdx)$ such that

$$f^\varepsilon(t)\in\mathcal{K}(T),\quad (t,\varepsilon)\in[0,T]\times(0,\varepsilon(T)). \tag{41}$$

We next study the equicontinuity of (f^ε) with respect to time and claim that

$$\lim_{h\to0}\sup_{t\in[0,T-h]}\sup_{\varepsilon\in(0,\varepsilon(T))}\left|\int_0^\infty\left(f^\varepsilon(t+h,x)-f^\varepsilon(t,x)\right)\ \varphi(x)\ \min\left(x,1\right)\ dx\right|=0 \tag{42}$$

for $\varphi\in W^{1,\infty}(0,+\infty)$. Indeed, consider $\varepsilon\in(0,\varepsilon(T))$, $h\in(0,T)$, and $t\in(0,T-h)$. By (23) we have

$$\begin{aligned}&\left|\int_0^\infty\left(f^\varepsilon(t+h,x)-f^\varepsilon(t,x)\right)\ \varphi(x)\ \min\left(x,1\right)\ dx\right|\\ &\le|\varphi_x|_{L^\infty}\int_t^{t+h}\int_0^\infty|\mathcal{V}^\varepsilon(s,x)|\ f^\varepsilon(s,x)\ \min\left(x,1\right)\ dxds\\ &\quad+|\varphi|_{L^\infty}\int_t^{t+h}\int_0^1|\mathcal{V}^\varepsilon(s,x)|\ f^\varepsilon(s,x)\ dxds.\end{aligned}$$

Now (10)–(11) and Lemma 6 entail that $|\mathcal{V}^\varepsilon(s,x)| \le C(T)\ (1+x)$ for $(s,x) \in (0,T)\times\mathbb{R}_+$. Consequently, thanks to (25) and (26), we have

$$\left|\int_0^\infty (f^\varepsilon(t+h,x)-f^\varepsilon(t,x))\ \varphi(x)\ \min(x,1)\ dx\right|$$
$$\le C(T)\ |\varphi|_{W^{1,\infty}} \int_t^{t+h}\int_0^\infty (1+x)\ f^\varepsilon(s,x)\ dxds \le C(T)\ |\varphi|_{W^{1,\infty}}\ h,$$

from which the claim (42) follows. Furthermore, since an arbitrary function φ in $L^\infty(\mathbb{R}_+)$ is the almost everywhere limit of a sequence of functions in $W^{1,\infty}(\mathbb{R}_+)$ which is bounded in $L^\infty(\mathbb{R}_+)$, it follows from (41) and (42) that (42) is actually valid for every $\varphi \in L^\infty(\mathbb{R}_+)$. We have thus proved that

$$\text{(43)}\qquad \begin{cases}\text{the family } \{f^\varepsilon,\ \ \varepsilon\in(0,\varepsilon(T))\} \text{ is weakly equicontinuous in}\\ L^1(\mathbb{R}_+;\min(x,1)dx) \text{ at every } t\in[0,T] \text{ (see [10, Definition 1.3.1])}.\end{cases}$$

Now, according to a variant of the Arzelà–Ascoli theorem (see, e.g., [10, Theorem 1.3.2]), we infer from (41) and (43) that

$$(f^\varepsilon) \text{ is relatively compact in } \mathcal{C}([0,T];w-L^1(\mathbb{R}_+;\min(x,1)dx)).$$

Once more using (41), we actually obtain that (f^ε) is relatively compact in $\mathcal{C}([0,T];w-L^1(\mathbb{R}_+;xdx))$. This last fact and Lemma 6 yield that there are a sequence (ε_n), $\varepsilon_n\to 0$, and functions

$$f\in\mathcal{C}([0,T];w-L^1(\mathbb{R}_+;xdx)) \quad\text{and}\quad u\in L^\infty(0,T)$$

such that

$$\text{(44)}\qquad \lim_{n\to+\infty}\sup_{t\in[0,T]}\left|\int_0^\infty (f^{\varepsilon_n}(t,x)-f(t,x))\ \varphi(x)\ x\ dx\right| = 0,$$

$$\text{(45)}\qquad u^{\varepsilon_n}\overset{*}{\rightharpoonup} u \text{ in } L^\infty(0,T)$$

for every $\varphi\in L^\infty(\mathbb{R}_+)$. Let $t\in[0,T]$. As $f^{\varepsilon_n}(t)$ is nonnegative a.e. in $\mathbb{R}_+$, a first consequence of (44) is that $f(t)$ is nonnegative a.e. in $\mathbb{R}_+$. Similarly we deduce from (45) that u is nonnegative a.e. in $(0,T)$. It also readily follows from (44) that $f(0)=f_0$ and

$$\lim_{n\to+\infty}\int_0^\infty x\ f^{\varepsilon_n}(t,x)\ dx = \int_0^\infty x\ f(t,x)\ dx$$

for $t\in[0,T]$. We may then pass to the limit in (25) and use (45) to obtain that

$$\int_0^\infty x\ f(t,x)\ dx = \int_0^\infty x\ f_0(x)\ dx, \quad t\in[0,T].$$

Owing to (44) and (45), we may also pass to the limit in (23) to obtain that (f,u) satisfies (15) with $\mathcal{V}$ given by (16). Owing to (9)–(11) and the integrability properties of f, we may take $g(x)=x$ as a test function in (15) and deduce that u satisfies (17). Also, proceeding as in [3] yields that $f\in\mathcal{C}([0,T];L^1(\mathbb{R}_+;xdx))$.

Finally, another consequence of (44) and (26) is that

$$\lim_{n\to+\infty}\int_\delta^\infty f^{\varepsilon_n}(t,x)\ dx = \int_\delta^\infty f(t,x)\ dx \le \int_0^\infty f_0(x)\ dx,$$

and the Fatou lemma guarantees that

$$\int_0^\infty f(t,x)\ dx \le \int_0^\infty f_0(x)\ dx, \quad t \in [0,T].$$

The above analysis being valid for an arbitrary $T \in \mathbb{R}_+$, we may take T to be an arbitrary large integer and perform countably many successive extractions to complete the proof of Theorem 1. □

We conclude this section with a remark on the almost everywhere finiteness of some negative moments of solutions to (3)–(5).

PROPOSITION 9. *Let f_0, k, and q be functions enjoying the properties* (9)–(12) *and denote by f a solution to* (3)–(5) *in the sense of Theorem* 1. *For $\alpha \in (0,1)$ and $T \in \mathbb{R}_+$ there holds*

$$\int_0^T \int_0^\infty x^{\alpha-1}\ q(x)\ f(t,x)\ dxdt < \infty.$$

Proof. Consider $\delta \in (0,1)$ and put $g_\delta(x) = (x+\delta)^\alpha - \delta^\alpha$ for $x \in \mathbb{R}_+$. Owing to (14), we may take g_δ as a test function in (15) and use the nonnegativity of f and u to obtain

$$\begin{aligned}
&\alpha \int_0^T \int_0^\infty (x+\delta)^{\alpha-1}\ q(x)\ f(t,x)\ dxdt \\
&\le \int_0^\infty x^\alpha\ f_0(x)\ dx + \alpha \int_0^T \int_0^\infty (x+\delta)^{\alpha-1}\ k(x)\ u(t)\ f(t,x)\ dxdt \\
&\le \int_0^\infty (1+x)\ f_0(x)\ dx + \alpha\ |u|_{L^\infty(0,T)} \int_0^T \int_0^\infty (x+\delta)^{\alpha-1}\ k(x)\ f(t,x)\ dxdt.
\end{aligned}$$

Recalling (13), there is x_T such that

$$q(x) \ge 2\ |u|_{L^\infty(0,T)}\ k(x)$$

for $x \in (0, x_T)$. Consequently,

$$\begin{aligned}
&\alpha \int_0^T \int_0^\infty (x+\delta)^{\alpha-1}\ q(x)\ f(t,x)\ dxdt \\
&\le C + \frac{\alpha}{2} \int_0^T \int_0^{x_T} (x+\delta)^{\alpha-1}\ q(x)\ f(t,x)\ dxdt \\
&\quad + \alpha\ x_T^{\alpha-1}\ |u|_{L^\infty(0,T)} \int_0^T \int_{x_T}^\infty k(x)\ f(t,x)\ dxdt \\
&\le C(T,\alpha) + \frac{\alpha}{2} \int_0^T \int_0^\infty (x+\delta)^{\alpha-1}\ q(x)\ f(t,x)\ dxdt,
\end{aligned}$$

where we have used (10) and (14) to obtain the last estimate. Therefore,

$$\int_0^T \int_0^\infty (x+\delta)^{\alpha-1}\ q(x)\ f(t,x)\ dxdt \le C(T,\alpha)$$

and we may let $\delta \to 0$ and use the Fatou lemma to complete the proof of Proposition 9. □

Remark. When k and q are given by (6) and $\alpha = 1/3$, Proposition 9 is similar to the last assertion of [7, Corollary 2.5] but stated in a different way.

4. Uniqueness. Throughout this section, f_0 and $\hat{f}_0$ are two functions satisfying (9), while k and q are two functions enjoying the properties (10)–(12) and (19) as well. Let (f,u) and $(\hat{f},\hat{u})$ be two solutions to (3)–(5) in the sense of Theorem 1 with data (f_0,k,q) and $(\hat{f}_0,k,q)$, respectively, with the obvious notation

$$\mathcal{V}(t,x)=k(x)\ u(t)-q(x),\quad \hat{\mathcal{V}}(t,x)=k(x)\ \hat{u}(t)-q(x),\quad (t,x)\in[0,+\infty)\times\mathbb{R}_+.$$

Next we introduce

$$\begin{aligned}&F(t,x)=\int_x^\infty f(t,y)\ dy,\quad \hat{F}(t,x)=\int_x^\infty \hat{f}(t,y)\ dy,\\ &E(t,x)=F(t,x)-\hat{F}(t,x)\end{aligned}$$

for $(t,x)\in[0,+\infty)\times\mathbb{R}_+$. We infer from (14) that

$$F\in W^{1,1}(0,T;L^1(\mathbb{R}_+))\quad\text{with}\quad F_x=-f\in L^\infty(0,T;L^1(\mathbb{R}_+;xdx))$$

and $F_t=\mathcal{V}\ f$ for each $T\in(0,+\infty)$. The function $\hat{F}$ enjoying similar properties, we conclude that, for each $T\in(0,+\infty)$,

$$E\in W^{1,1}(0,T;L^1(\mathbb{R}_+))\quad\text{with}\quad E_x=\hat{f}-f\in L^\infty(0,T;L^1(\mathbb{R}_+;xdx))$$

and

$$E_t=\mathcal{V}\ f-\hat{\mathcal{V}}\ \hat{f}=-\mathcal{V}\ E_x+\left(\mathcal{V}-\hat{\mathcal{V}}\right)\ \hat{f}. \tag{46}$$

Since $k(0)=0$ by (10), we infer from (46) that

$$\begin{aligned}\int_0^\infty |E(t,x)|\ dx\le{}&\int_0^\infty |E(0,x)|\ dx-q(0)\ \int_0^t |E(s,0)|\ ds\\ &+\int_0^t\int_0^\infty \mathcal{V}_x(s,x)\ |E(s,x)|\ dxds\\ &+\int_0^t\int_0^\infty k(x)\ \hat{f}(s,x)\ \left|(u-\hat{u})(s)\right|\ dxds.\end{aligned} \tag{47}$$

A formal proof of (47) follows by multiplying the equation satisfied by E by $sign(E)$ and integrating by parts. We next compute the last term of the right-hand side of (47) as follows. Since $\hat{f}$ and k are nonnegative, we have

$$\begin{aligned}&\int_0^\infty k(x)\ \hat{f}(s,x)\ \left|(u-\hat{u})(s)\right|\ dx\\ &=\left|u(s)\ \int_0^\infty k(x)\ \hat{f}(s,x)\ dx-\int_0^\infty q(x)\ \hat{f}(s,x)\ dx\right|\\ &=\left|u(s)\ \int_0^\infty k(x)\ \left(\hat{f}-f\right)(s,x)\ dx-\int_0^\infty q(x)\ \left(\hat{f}-f\right)(s,x)\ dx\right|\\ &=\left|u(s)\ \int_0^\infty k(x)\ E_x(s,x)\ dx-\int_0^\infty q(x)\ E_x(s,x)\ dx\right|.\end{aligned}$$

Now it readily follows from (10), (11), and (14) that

$$\begin{aligned}&\int_0^\infty k(x)\ E_x(s,x)\ dx=-\int_0^\infty k'(x)\ E(s,x)\ dx,\\ &-\int_0^\infty q(x)\ E_x(s,x)\ dx=q(0)\ E(s,0)+\int_0^\infty q'(x)\ E(s,x)\ dx.\end{aligned}$$

We thus end up with

$$\int_0^\infty k(x)\ \hat{f}(s,x)\ \left|(u-\hat{u})(s)\right|\ dx = \left| q(0)\ E(s,0) - \int_0^\infty \mathcal{V}_x(s,x)\ E(s,x)\ dx \right|.$$

After inserting the above formula in (47), we are led to

$$\begin{aligned}\int_0^\infty |E(t,x)|\ dx \le \int_0^\infty |E(0,x)|\ dx - q(0)\ \int_0^t |E(s,0)|\ ds \\ + \int_0^t \int_0^\infty \mathcal{V}_x(s,x)\ |E(s,x)|\ dxds \\ + q(0)\ \int_0^t |E(s,0)|\ ds + \int_0^t \int_0^\infty |\mathcal{V}_x(s,x)|\ |E(s,x)|\ dxds,\end{aligned}$$

whence

$$\begin{aligned}\int_0^\infty |E(t,x)|\ dx \le \int_0^\infty |E(0,x)|\ dx \\ + \int_0^t \int_0^\infty \left(|\mathcal{V}_x(s,x)| + \mathcal{V}_x(s,x)\right)\ |E(s,x)|\ dxds.\end{aligned} \tag{48}$$

We finally consider $T \in \mathbb{R}_+$ and $t \in [0,T]$. On the one hand, it follows from (14) that $u \in L^\infty(0,T)$. On the other hand, notice that

$$|\mathcal{V}_x(s,x)| + \mathcal{V}_x(s,x) = \begin{cases} 0 & \text{if} \quad \mathcal{V}_x(s,x) \le 0, \\ \\ 2\ \mathcal{V}_x(s,x) & \text{otherwise,} \end{cases}$$

and

$$\mathcal{V}_x(s,x) \le \left(|u|_{L^\infty(0,T)}\ k'(x) - q'(x)\right)$$

in the latter case. The condition (19) then warrants that $(s,x) \to |\mathcal{V}_x(s,x)| + \mathcal{V}_x(s,x)$ belongs to $L^\infty((0,T)\times\mathbb{R}_+)$. We may now apply the Gronwall lemma to (48) and obtain that there is a positive constant γ (depending on T, k, q, and u) such that

$$\int_0^\infty |E(t,x)|\ dx \le \left(\int_0^\infty |E(0,x)|\ dx\right)\ e^{\gamma t}$$

for each $t \in [0,T]$. Theorem 2 then readily follows by taking $f_0 = \hat{f}_0$.

REFERENCES

[1] J.F. Collet and T. Goudon, *On solutions of the Lifshitz-Slyozov model*, Nonlinearity, 13 (2000), pp. 1239–1262.

[2] C. Dellacherie and P.A. Meyer, *Probabilités et potentiel, chapitres* I *à* IV, Hermann, Paris, 1975.

[3] R.J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[4] Ph. Laurençot, *Weak solutions to the Lifshitz-Slyozov-Wagner equation*, Indiana Univ. Math. J., 50 (2001), pp. 1319–1346.

[5] Lê Châu-Hoàn, *Etude de la classe des opérateurs m-accrétifs de* $L^1(\Omega)$ *et accrétifs dans* $L^\infty(\Omega)$, Thèse de 3ème cycle, Université de Paris VI, Paris, 1977.

[6] I.M. Lifshitz and V.V. Slyozov, *The kinetics of precipitation from supersaturated solid solutions*, J. Phys. Chem. Solids, 19 (1961), pp. 35–50.
[7] B. Niethammer and R.L. Pego, *On the initial-value problem in the Lifshitz–Slyozov–Wagner theory of Ostwald ripening*, SIAM J. Math. Anal., 31 (2000), pp. 467–485.
[8] B. Niethammer and R.L. Pego, in preparation.
[9] V.V. Slezov and V.V. Sagalovich, *Diffusive decomposition of solid solutions*, Soviet Phys. Uspekhi, 30 (1987), pp. 23–45.
[10] I.I. Vrabie, *Compactness Methods for Nonlinear Evolutions*, 2nd ed., Pitman Monogr. Surveys Pure Appl. Math. 75, Longman, Harlow, UK, 1995.
[11] C. Wagner, *Theorie der Alterung von Niederschlägen durch Umlösen (Ostwald-Reifung)*, Z. Elektrochem., 65 (1961), pp. 581–591.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 273–291

ASYMPTOTIC BEHAVIOR OF A ONE-DIMENSIONAL COMPRESSIBLE VISCOUS GAS WITH FREE BOUNDARY*

TAO PAN†, HONGXIA LIU‡, AND KENJI NISHIHARA§

Abstract. We consider the initial-boundary value problem for a one-dimensional compressible viscous gas with free boundary, which is modeled in the Eulerian coordinate as

$$\text{(IBVP)}\qquad \begin{cases} \rho_t + (\rho u)_x = 0, \quad x > x(t), \quad t > 0, \\ (\rho u)_t + (\rho u^2 + p)_x = \mu u_{xx}, \quad x > x(t), \quad t > 0, \\ (p - \mu u_x)|_{x=x(t)} = p_0, \quad \frac{dx(t)}{dt} = u(x(t), t), \quad t \geq 0, \\ (\rho, u)|_{t=0} = (\rho_0, u_0)(x), \quad x \geq x(0). \end{cases}$$

Here, $\rho\,(> 0)$ is the density, u is the velocity, $p = p(\rho) = \rho^\gamma$ ($\gamma \geq 1$: the adiabatic constant) is the pressure, and $\mu\,(> 0)$ is the viscosity constant. At the boundary the flow is attached to the atmosphere with pressure $p_0\,(> 0)$ and the boundary condition is derived by the balance law. The initial data have constant states (ρ_+, u_+) at $x = \infty$. The flow has no vacuum state so that $\rho_0(x) > 0$ and $\rho_+ > 0$ are assumed. Our main purpose is to investigate the asymptotic behaviors of solutions for (IBVP), which are closely related to those for the corresponding Cauchy problem and hence the corresponding Riemann problem. Depending on p_0 and the endstates (ρ_+, u_+), the solutions are shown to tend to the outgoing rarefaction wave or the outgoing viscous shock wave as t tends to infinity. The proof is given under the weakness assumption of the waves. The analysis will be done by changing (IBVP) into the problem in the Lagrangian coordinate.

Key words. one-dimensional compressible viscous gas, free boundary, asymptotic behavior, rarefaction wave, viscous shock wave

AMS subject classification. 35L65

PII. S0036141001385745

1. Introduction.

A one-dimensional barotropic viscous flow is modeled in the Eulerian coordinate $(\tilde{x}, \tilde{t})$ as

$$\text{(1.1)}\qquad \begin{cases} \tilde{\rho}_{\tilde{t}} + (\tilde{\rho}\tilde{u})_{\tilde{x}} = 0, \\ (\tilde{\rho}\tilde{u})_{\tilde{t}} + (\tilde{\rho}\tilde{u}^2 + \tilde{p})_{\tilde{x}} = \mu \tilde{u}_{\tilde{x}\tilde{x}}, \end{cases}$$

where $\tilde{\rho}$ is the density, $\tilde{u}$ is the velocity, $\tilde{p} = \tilde{p}(\tilde{\rho}) = \tilde{\rho}^\gamma$ ($\gamma \geq 1$: the adiabatic constant) is the pressure, and $\mu\,(> 0)$ is the viscosity constant. If the flow is attached at the boundary to the atmosphere with pressure p_0, then the balance law at the boundary $\tilde{x} = \tilde{x}(\tilde{t})$ gives the condition

$$\text{(1.2)}\qquad (\tilde{p} - \mu \tilde{u}_{\tilde{x}})|_{\tilde{x}=\tilde{x}(\tilde{t})} = p_0 \quad \text{and} \quad \frac{d\tilde{x}(\tilde{t})}{d\tilde{t}} = \tilde{u}(\tilde{x}(\tilde{t}), \tilde{t})$$

*Received by the editors March 1, 2001; accepted for publication (in revised form) February 27, 2002; published electronically October 8, 2002. The research of the first and second authors was supported by the National Natural Science Foundation of China (grant 10061001) and the Natural Science Foundation of Guangxi (grants 9912020 and 0135001).

http://www.siam.org/journals/sima/34-2/38574.html

†Department of Mathematics and Information Sciences, Guangxi University, Nanning, 530004, People's Republic of China (tpan2000@hotmail.com), and Laboratory of Bioinformation and Food Engineering, Faculty of Bioresources, Mie University, Tsu, Mie, 514-8507, Japan (pan@bife.bio.mie-u.ac.jp).

‡Department of Mathematics, Jinan University, Guangzhou, 510632, People's Republic of China (hongxia-liu@163.net).

§School of Political Science and Economics, Waseda University, Tokyo 169-8050, Japan (kenji@mn.waseda.ac.jp). The research of this author was supported in part by Grant-in-Aid for Scientific Research (grant c(2)10640216) of the Ministry of Education, Science, Sports and Culture.

by $(1.1)_2$ (second equation of (1.1)), which implies that $\tilde{x} = \tilde{x}(\tilde{t})$ is the free boundary.

We now consider (1.1) on $\tilde{x} > \tilde{x}(\tilde{t})$ with the boundary condition (1.2) and the initial data

$$(\tilde{\rho}, \tilde{u})|_{\tilde{t}=0} = (\tilde{\rho}_0, \tilde{u}_0)(\tilde{x}), \quad \tilde{x} \geq \tilde{x}(0) =: 0. \tag{1.3}$$

The initial data are assumed to be constant as $\tilde{x} \to \infty$:

$$\lim_{\tilde{x}\to\infty} (\tilde{\rho}_0, \tilde{u}_0)(\tilde{x}) = (\rho_+, u_+). \tag{1.4}$$

Also,

$$0 < \tilde{\rho}_0(\tilde{x}) < \infty, \quad \rho_+ > 0, \quad \text{and} \quad p_0 > 0 \tag{1.5}$$

are assumed, so that the flow has no vacuum state.

Our main interest concerns the large-time behaviors of solutions to (1.1)–(1.3). To explore those, we transform the Eulerian coordinate $(\tilde{x}, \tilde{t})$ into the Lagrangian coordinate (x, t) by

$$x = \int_0^{\tilde{x}} \tilde{\rho}_0(y)dy, \quad \tilde{t} = t.$$

Then, (1.1)–(1.3) changes into the problem with fixed boundary in the form of

$$\begin{cases} v_t - u_x = 0, & x \in \mathbf{R}_+ = (0, \infty), \quad t > 0, \\ u_t + p(v)_x = \mu \left(\dfrac{u_x}{v}\right)_x \end{cases} \tag{1.6}$$

with the boundary condition

$$\left(p(v) - \mu \frac{u_x}{v}\right)(0, t) = p_0, \quad t \geq 0, \tag{1.7}$$

and the initial condition

$$(v, u)(x, 0) = (v_0, u_0)(x), \quad x \in \mathbf{R}_+, \tag{1.8}$$

where $\tilde{u}(\tilde{x}, \tilde{t}) = u(x, t)$, etc., and $v = 1/\rho$, so that $p(v) := \tilde{p}(\tilde{\rho}) = v^{-\gamma}$ satisfies

$$p'(v) < 0, \quad p''(v) > 0 \quad \text{for} \quad v > 0. \tag{1.9}$$

The assumptions (1.4) and (1.5) are written as

$$\lim_{x\to\infty} (v_0(x), u_0)(x) = (v_+, u_+), \quad v_+ = 1/\rho_+ \tag{1.10}$$

and

$$0 < v_0(x) < \infty, \quad v_+ > 0, \quad \text{and} \quad p_0 > 0. \tag{1.11}$$

The cases of the Dirichlet boundary

$$u|_{x=0} = u_- := 0 \tag{1.12}$$

instead of (1.7) have been investigated by Matsumura and Mei [3], Pan, Liu, and Nishihara [11], and Matsumura and Nishihara [8]. From those results the behaviors

of solutions closely relate to those for the corresponding Cauchy problem on $\mathbf{R} = (-\infty, \infty)$ for (1.6) with $(v,u)|_{t=0} = (v_0,u_0)(x) \to (v_\pm, u_\pm)$ as $x \to \infty$, and hence the corresponding Riemann problem for (1.6) with $\mu = 0$ for the Riemann data

$$(v_0^R, u_0^R)(x) = \begin{cases} (v_-, u_-), & x < 0, \\ (v_+, u_+), & x > 0. \end{cases}$$

As is well known, the behaviors expected for the Cauchy problem and the Riemann problem divide $\mathbf{R}^2_{(v,u)}$-space into four ranges by the shock curves $S_i(v_-, u_-)$ and the rarefaction curves $R_i(v_-, u_-)$ $(i = 1, 2)$. Refer to [5, 7, 12] and the survey paper [10].

Roughly speaking, it is shown in [3] that if $u_- > u_+$, then there is a constant v_- such that $(v_+, u_+) \in S_2(v_-, u_-)$ (the 2-shock curve) and the solution (v,u) to (1.6), (1.12), (1.8) tends to the 2-viscous shock wave $(V_2, U_2)(x - st + \alpha)$ connecting (v_-, u_-) and (v_+, u_+) as $t \to \infty$ for some shift α determined by the initial data. On the other hand, in [11, 8] they have shown that if $u_- < u_+$, then there is a constant v_- such that $(v_+, u_+) \in R_2(v_-, u_-)$ (the 2-rarefaction curve) and the solution (v,u) tends to the 2-rarefaction wave $(v_2^r, u_2^r)(x/t)$ connecting (v_-, u_-) and (v_+, u_+) as $t \to \infty$.

Summing up their results, all waves are reflected at the boundary and they merge to the outgoing wave, which is the 2-viscous shock wave or the 2-rarefaction wave depending on the data u_- and (v_+, u_+).

Thus, in our problem it is also a key point to determine the value $v_- = v(0,t)$ instead of (1.12). From $(1.6)_1$, the boundary condition (1.7) can be rewritten as

$$\left(p(v) - \mu \frac{v_t}{v}\right)(0,t) = p_0. \tag{1.13}$$

If we define v_- by

$$p(v_-) = p_0 \quad \text{or} \quad v_- = p_0^{1/\gamma}, \tag{1.14}$$

then (1.13) becomes the ordinary differential equation

$$\mu \frac{v_t}{v}(0,t) = p(v(0,t)) - p(v_-). \tag{1.15}$$

The initial data of $v(0,t)$ should be

$$v(0,t)|_{t=0} = v_0(0) \tag{1.16}$$

from the compatibility condition. Solving (1.15)–(1.16) we have

$$\lim_{t \to \infty} v(0,t) = v_-, \tag{1.17}$$

which will be shown in the next section.

Thus, we can expect that if $v_- > v_+$, then u_- is uniquely determined by $(v_+, u_+) \in R_2(v_-, u_-)$ and the solution (v,u) to (1.6)–(1.8) tends to the 2-rarefaction wave $(v_2^r, u_2^r)(x/t)$ connecting (v_-, u_-) and (v_+, u_+) as $t \to \infty$, and that if $v_- < v_+$, then u_- is uniquely determined by $(v_+, u_+) \in S_2(v_-, u_-)$ and the solution (v,u) to (1.6)–(1.8) tends to the 2-viscous shock wave $(V_2, U_2)(x - s_2 t + \alpha)$ as $t \to \infty$ for suitable α. In the results, in each case $u(0,t) \to u_-$ as $t \to \infty$, and hence the free boundary $\tilde{x}(\tilde{t})$ in the Eulerian coordinate satisfies $\frac{d\tilde{x}(\tilde{t})}{d\tilde{t}} = \tilde{u}(\tilde{x}(\tilde{t}), \tilde{t}) \to u_-$ as $\tilde{t} \to \infty$.

Our purpose in this paper is to show these two assertions under some weakness condition, that is, $|v_+ - v_-|$ is suitably small. We note that the weakness conditions

are assumed in [3] and [11]. In [8] any smallness conditions are not assumed. A global result corresponding to [8] will be expected in our problem, which will be investigated in a forthcoming paper.

Related to the boundary effect, we mention about another kind of initial and boundary problems, i.e., the inflow problem and outflow problem. Those problems have been recently proposed by Matsumura, and all behaviors of solutions expected are classified in [2]. In [9] some cases are proved rigorously. Other cases remain open.

The outline of this paper is as follows. In the next section the behavior of $v(0,t)$ will be studied. In section 3 the convergence to the 2-rarefaction wave will be treated, and the convergence to the 2-viscous shock wave will be done in the final section.

2. Preliminaries. We observe the behaviors of the boundary value $v(0,t)$. Denote $v(t) := v(0,t)$ in this section; then $v(t)$ satisfies the ordinary differential equation

$$\begin{cases} \frac{dv}{dt} = \frac{1}{\mu}\left(p\left(v\right) - p\left(v_-\right)\right)v, \\ v\left(0\right) = v_0(0) =: v_0 \end{cases} \tag{2.1}$$

by (1.15)–(1.16). Note that v_- is defined by (1.14) and $v_0(x)$ is a initial value in (1.8).

LEMMA 2.1. *Under the condition* (1.9)*, the global smooth solution* $v = v\left(t\right)$ *for* (2.1) *satisfies the following properties:*

(i) *If* $v_0 = v_-$*, then* $v\left(t\right) \equiv v_-$*. If* $v_0 \neq v_-$*, then* $v(t) \neq v_-$ *for any* $t \in \mathbf{R}_+$*.*

(ii) *If* $v_0 > v_-$*, then* $v_- < v\left(t\right) < v_0$*,* $v'\left(t\right) < 0$*,* $v''\left(t\right) > 0$*.*

(iii) *If* $v_0 < v_-$*, then* $v_0 < v\left(t\right) < v_-$*,* $v'\left(t\right) > 0$*,* $v''\left(t\right) < 0$*.*

(iv) $\lim_{t\to+\infty} v\left(t\right) = v_-$*.*

Proof. (i) If there is a global smooth solution for (2.1), then it is easy to show that the solution is unique. Hence $v(t) \equiv v_-$ if $v_0 = v_-$. If $v_0 > v_-$ and $v(t_0) = v_-$ with $v(t) > v_-$ for $0 \le t < t_0 < \infty$, then

$$\int_0^t \frac{\frac{dv}{d\tau}(\tau)}{v(\tau) - v_-} d\tau = \int_0^t \frac{1}{\mu}\frac{p(v(\tau)) - p(v_-)}{v(\tau) - v_-} v(\tau) d\tau, \quad t < t_0.$$

When $\tau \to t_0 - 0$, the left-hand side tends to $-\infty$, and the right-hand side is larger than

$$-\frac{t_0}{\mu}\max_{[0,t_0]}\left\{\frac{p(v(\tau)) - p(v_-)}{v(\tau) - v_-} v(\tau)\right\} \ (> -\infty),$$

which deduces the contradiction. Hence we obtain the property (i).

(ii) Rewrite (2.1) as

$$\int_{v_0}^{v} \frac{dw}{\left(p\left(w\right) - p\left(v_-\right)\right)w} = \int_0^t \frac{dt}{\mu} = \frac{t}{\mu}. \tag{2.2}$$

By the mean-value theorem

$$\int_{v_0}^{v} \frac{dw}{\left(w - v_-\right)wp'\left(\xi\right)} = \frac{t}{\mu}, \tag{2.3}$$

where ξ is between v_0 and v.

Suppose that there exists $t_0 > 0$ such that $v\left(t_0\right) = \bar{v} \ge v_0 > v_-$; then from (2.3), we have

$$\int_{v_0}^{\bar{v}} \frac{dw}{\left(w - v_-\right)wp'\left(\xi\right)} = \frac{t_0}{\mu}.$$

The left-hand side of the above equality is negative, in which the signs contradict. Hence $v(t) < v_0$. From $v(0) = v_0 \neq v_-$, $v(t) \not\equiv v_-$ for $t \geq 0$. If $v(t) \geq v_-$ is not valid, then there exist $0 < t_0 < t_1$ such that $v(t_0) = v_- > v(t_1)$. Therefore, as $t \geq t_0$, $v(t) \equiv v_-$ is the solution of (2.1). This contradicts $v(t_1) < v_-$. Thus, we have proved that $v_- < v(t) < v_0$. Hence $v'(t) = \frac{1}{\mu}(p(v) - p(v_-))v < 0$ and $v''(t) = \frac{1}{\mu}(p'(v)v + p(v) - p(v_-))v'(t) > 0$.

(iii) The proof is similar to that of (ii).

(iv) We show $\lim_{t\to+\infty} v(t) = v_-$ when $v_0 > v_-$. If not, then there exists $\varepsilon > 0$ such that for any $t > 0$, there is a positive constant $t_1 > t$ such that $v(t_1) \geq v_- + \varepsilon$. Since $(p(v) - p(v_-))v$ is decreasing for $v > v_-$,

$$v'(t_1) = (p(v(t_1)) - p(v_-))\frac{v(t_1)}{\mu} \leq \frac{(v_- + \varepsilon)}{\mu}(p(v_- + \varepsilon) - p(v_-)) \triangleq -\nu.$$

In view of $v''(t) > 0$, we have $v'(t) \leq v'(t_1) \leq -\nu$. Let $t \to +\infty$; then $v(t) \to -\infty$. This generates a contradiction. Thus $\lim_{t\to+\infty} v(t) = v_-$. The proof in the case of $v_0 < v_-$ is similar.

Conversely, we show the global existence and precise behavior for $v(t)$. □

LEMMA 2.2. *Suppose that* (1.9) *holds. Then there exists a unique global smooth solution* $v(t)$ *to* (2.1). *Moreover,* $|v(t) - v_-| = O(1)|v_0 - v_-|e^{-c_0 t}$ *and* $|v'(t)| = O(1)|v_0 - v_-|e^{-c_0 t}$ *as* $t \to +\infty$, *where* $c_0 = \frac{1}{\mu}v_-|p'(v_-)|$.

Proof. The local existence of a smooth solution to (2.1) can be shown in the standard way. Making use of the local existence result and the a priori estimate in Lemma 2.1, we can prove the existence of the unique global smooth solution to (2.1) through the continuation process.

Next, we shall prove $v(t) - v_-$, $v'(t) \to 0$ exponentially as $t \to +\infty$ in the case of $v_0 > v_-$. By Lemma 2.1 $v_- < v(t) < v_0$. By using the Taylor expansion theorem, we can rewrite (2.2) as

$$\int_v^{v_0} \frac{dw}{(w - v_-)\left(1 - \frac{p''(\eta)}{2|p'(v_-)|}(w - v_-)\right)w|p'(v_-)|} = \frac{t}{\mu}, \tag{2.4}$$

where $\eta = v_- + \theta(w - v_-)$, $0 < \theta < 1$.

Put $m_0 = \sup_{v_-\leq v\leq v_0} \frac{p''(\eta)}{2|p'(v_-)|}$ and take $v_1 = v(t_1) > v_-$, $t_1 \gg 1$ such that $1 - m_0(v_1 - v_-) \geq \frac{1}{2}$, and then (2.4) is rewritten as

$$\left(\int_v^{v_1} + \int_{v_1}^{v_0}\right)\frac{v_- dw}{(w - v_-)(1 - m_0(w - v_-))w} \geq \frac{1}{\mu}v_-|p'(v_-)|t = c_0 t,$$

and hence

$$\int_v^{v_1} \frac{v_- dw}{(w - v_-)(1 - m_0(w - v_-))w} \geq c_0 t - c(v_0, v_1),$$

where $c(v_0, v_1) = \int_{v_1}^{v_0} \frac{v_- dw}{(w-v_-)(1-m_0(w-v_-))w}$.

Using the equality

$$\frac{v_-}{(w - v_-)(1 - m_0(w - v_-))w} = \frac{1}{w - v_-} + \frac{B_0 m_0}{1 - m_0(w - v_-)} - \frac{A_0}{w},$$

where $A_0 = \frac{1}{1+m_0 v_-}$, $B_0 = \frac{m_0 v_-}{1+m_0 v_-}$, we obtain

$$\ln \frac{v - v_-}{(1 - m_0 (v - v_-))^{B_0} v^{A_0}} \Bigg|_v^{v_1} \geq c_0 t - c (v_0, v_1),$$

which gives

$$\begin{aligned} v - v_- &\leq (v_1 - v_-) \frac{(1-m_0(v-v_-))^{B_0} v^{A_0}}{(1-m_0(v_1-v_-))^{B_0} v_-^{A_0}} e^{-c(v_0,v_1)} e^{-c_0 t} \\ &\leq \frac{v_0^{A_0}}{(1-m_0(v_1-v_-))^{B_0} v_-^{A_0}} e^{-c(v_0,v_1)} |v_0 - v_-| e^{-c_0 t}. \end{aligned} \tag{2.5}$$

On the other hand, by (2.4),

$$\int_v^{v_0} \frac{v_- dw}{(w - v_-) w} \leq \frac{1}{\mu} v_- |p' (v_-)| t,$$

which implies that

$$\begin{aligned} v - v_- &\geq \frac{(v_0 - v_-) v}{v_0} e^{-c_0 t} \\ &\geq \frac{v_-}{v_0} |v_0 - v_-| e^{-c_0 t}. \end{aligned} \tag{2.6}$$

Thus we have the desired result $|v (t) - v_-| = O (1) |v_0 - v_-| e^{-c_0 t}$ by (2.5) and (2.6). From (2.1) we also have $|v' (t)| = O (1) |v_0 - v_-| e^{-c_0 t}$. □

3. Convergence to rarefaction waves. As stated in the introduction, the asymptotic behavior of the solution to (1.6)–(1.8) depends on the sign of $v_- - v_+$. This section is devoted to studying the case $v_- > v_+$, in which the solution to (1.6)–(1.8) converges to the 2-rarefaction waves.

3.1. Main result. To state our result, we first remember the results for the corresponding Riemann problem on $R = (-\infty, +\infty)$ for given constant states $(v_\pm, u_\pm)$, $v_\pm > 0$:

$$\begin{cases} v_t - u_x = 0, \\ u_t + p (v)_x = 0, \\ (v, u) (x, 0) = \left(v_0^R, u_0^R\right) (x) = \begin{cases} (v_-, u_-), & x < 0, \\ (v_+, u_+), & x > 0, \end{cases} \end{cases} \qquad t > 0, \quad x \in R. \tag{3.1}$$

It is well known that if $(v_+, u_+) \in R_i (v_-, u_-)$ $(i = 1, 2)$, then (3.1) admits an i-rarefaction wave solution

$$(v_i^r, u_i^r) (x/t) = \begin{cases} (v_-, u_-), & -\infty < \xi \leq \lambda_i (v_-), \\ \left(\lambda_i^{-1} (\xi), \ u_- - \int_{v_-}^{\lambda_i^{-1}(\xi)} \lambda_i (s) \, ds\right), & \lambda_i (v_-) \leq \xi \leq \lambda_i (v_+), \\ (v_+, u_+), & \lambda_i (v_+) \leq \xi < +\infty, \end{cases} \tag{3.2}$$

where $\lambda_i (v) = (-1)^i \sqrt{-p' (v)}$ $(i = 1, 2)$ are the eigenvalues for (3.1) and

$$R_i (v_-, u_-) = \left\{ (v, u) \in \Omega \,\middle|\, u = u_- - \int_{v_-}^{v} \lambda_i (s) \, ds, u \geq u_- \right\} \tag{3.3}$$

for a suitable neighborhood Ω of (v_-, u_-) in $R^2_{(v,u)}$.

Since there is a boundary at $x = 0$ in our problem, the backward flow reflects at the boundary and the total flow is eventually expected to move forward and behave as the 2-rarefaction wave for large time if $v_- > v_+$. Therefore, for any given $v_- = p^{-1}(p_0) > v_+ > 0$ and $u_+ \in R$, there is a unique $u_- \in R$ such that

$$(v_+, u_+) \in R_2(v_-, u_-) \tag{3.4}$$

and (3.1) admits the 2-rarefaction wave $(v_2^r, u_2^r)(x/t)$ connecting (v_-, u_-) and (v_+, u_+), and the solution (v, u) of (1.6)–(1.8) is expected to behave as $(v_2^r, u_2^r)(x/t)|_{x\geq 0}$.

We now assume

$$(v_0(\cdot) - v_+, u_0(\cdot) - u_+) \in L^2(R_+), \qquad (v_0(\cdot), u_0(\cdot))_x \in L^2(R_+) \tag{3.5}$$

and set

$$\begin{aligned}\Phi_0^2 = {} & \|(v_0(\cdot) - v_+, u_0(\cdot) - u_+)\|^2 + \|(v_0(\cdot), u_0(\cdot))_x\|^2 \\ & + |(v_+ - v_-, u_+ - u_-)| + |v_0(0) - v_-|.\end{aligned}$$

Then our first main theorem is the following.

THEOREM 3.1. *For given constants (v_+, u_+) and $v_- = p^{-1}(p_0)$ with $0 < v_+ < v_-$ and $u_- \in R$ determined by (3.4), there exists a positive constant ε such that if $\Phi_0 < \varepsilon$, then the initial-boundary value problem (1.6)–(1.8) has a unique global solution $(v, u)(x, t)$ in time satisfying*

$$\begin{cases} (v - v_+, u - u_+) \in C^0([0, +\infty); L^2), \\ (v, u)_x \in C^0([0, +\infty); L^2), \\ u_{xx} \in L^2([0, +\infty); L^2) \end{cases} \tag{3.6}$$

and

$$\lim_{t\to+\infty} \sup_{x\in R_+} |(v, u)(x, t) - (v_2^r, u_2^r)(x/t)| = 0, \tag{3.7}$$

where $(v_2^r, u_2^r)(x/t)$ is given by (3.2).

3.2. Smooth rarefaction wave. To prove Theorem 3.1 we start with the Riemann problem on $R = (-\infty, +\infty)$ for the typical Burgers equation

$$\begin{cases} \omega_t + \omega\omega_x = 0, & (x, t) \in R \times R_+, \\ \omega(0, x) = \omega_0^r(x) = \begin{cases} \omega_-, & x < 0, \\ \omega_+, & x > 0, \end{cases} \end{cases} \tag{3.8}$$

with $\omega_- < \omega_+$. As is well known, the problem (3.8) has the centered rarefaction wave $\omega(x, t) = \omega^r\left(\frac{x}{t}\right)$ given by

$$\omega^r\left(\frac{x}{t}\right) = \begin{cases} \omega_-, & x \leq \omega_- t, \\ \frac{x}{t}, & \omega_- t < x < \omega_+ t, \\ \omega_+, & x \geq \omega_+ t. \end{cases} \tag{3.9}$$

We approximate the rarefaction wave by the solution to the following problem:

$$\begin{cases} \omega_t + \omega\omega_x = 0, \\ \omega(0, x) = \omega_0(x) := \widehat{\omega} + \widetilde{\omega}\tanh x, \end{cases} \tag{3.10}$$

where $\widehat{\omega} = (\omega_+ + \omega_-)/2$, $\widetilde{\omega} = (\omega_+ - \omega_-)/2$.

Now, we state the properties of ω.

LEMMA 3.1. *Suppose $\omega_+ > \omega_- > 0$. Then (3.10) has a unique smooth global solution in time $\omega(x, t)$, which satisfies the following properties:*

(i) $\omega_- < \omega(x,t) < \omega_+, \quad \omega_x(x,t) > 0.$

(ii) *For any p $(1 \le p < +\infty)$, there exists a constant C_p such that for $t \ge 0$,*

$$\|\omega_x(\cdot,t)\|_{L^p} \le C_p \min\left\{\widetilde{\omega}, \widetilde{\omega}^{\frac{1}{p}} t^{-1+\frac{1}{p}}\right\},$$

$$\|\omega_{xx}(\cdot,t)\|_{L^p} \le C_p \min\left\{\widetilde{\omega}, t^{-1}\right\}.$$

(iii) *For any $x \le 0$ and $t \ge 0$,*

$$0 < \omega(x,t) - \omega_- \le \widetilde{\omega} \exp\left\{-2(|x| + \omega_- t)\right\},$$

$$0 < \omega_x(x,t) \le 2\widetilde{\omega} \exp\left\{-2(|x| + \omega_- t)\right\}.$$

(iv) $\lim_{t\to+\infty} \sup_{x\in R} |\omega(x,t) - \omega^r(x/t)| = 0.$

Since the 2-rarefaction waves $(v_2^r, u_2^r)(x,t)$ are constructed in (3.2), their smooth approximation $(V,U)(x,t)$ can be defined by

$$\lambda_2(V) = \omega(x,t), \qquad U = u_- - \int_{v_-}^{V(x,t)} \lambda_2(s)\, ds,$$

where ω is the solution to (3.10) with $\lambda_2(v_\pm) = \omega_\pm$. Then it can be easily seen that (V,U) is a smooth exact solution to $(3.1)_1$–$(3.1)_2$. Moreover, the properties of ω shift to those of (V,U).

LEMMA 3.2. *Let $\delta = |v_+ - v_-| + |u_+ - u_-|$. Then the following hold:*

(i) *There exists a positive constant C such that for any $t \ge 0$, $x \in R_+$, and $0 \le \delta \le \delta_0$,*

$$|V_x| \le CV_t, \qquad 0 < V_t = U_x \le C\delta,$$

$$|U_{xx}| \le C\left(|V_{xx}| + |V_x|^2\right).$$

(ii) *For any p $(1 \le p < +\infty)$, there exists a constant $C_p > 0$ such that for $t \ge 0$ and $0 \le \delta \le \delta_0$,*

$$\|(V_x, U_x)(\cdot,t)\|_{L^p(R_+)} \le C_p \delta^{\frac{1}{p}} (1+t)^{-1+\frac{1}{p}},$$

$$\|(V_{xx}, U_{xx})(\cdot,t)\|_{L^p(R_+)} \le C_p \min\left\{\delta, (1+t)^{-1}\right\}.$$

(iii) *There exist positive constants C and c_- depending only on $\lambda_2(v)$ and $v_\pm$ such that for $t \ge 0$ and $0 \le \delta \le \delta_0$,*

$$|V(0,t) - v_-, V_x(0,t), V_{xx}(0,t)| \le C\delta e^{-c_- t}.$$

(iv) $\lim_{t\to+\infty} \sup_{x\in R_+} \left|(V,U)(x,t) - (v_2^r, u_2^r)\left(\frac{x}{t}\right)\right| = 0.$

The proofs of Lemmas 3.1 and 3.2 are given in [6].

3.3. Reformulation of the problem. Rewrite (1.6)–(1.8) by the change of variables $(v,u) = (V+\phi, U+\psi)$ as follows:

$$\text{(3.11)} \quad \begin{cases} \phi_t - \psi_x = 0, \\ \psi_t + (p(V+\phi) - p(V))_x - \mu\left(\dfrac{\psi_x}{V+\phi}\right)_x = F, \quad F = \mu\left(\dfrac{U_x}{V+\phi}\right)_x, \\ \left(p(V+\phi) - \mu\dfrac{U_x+\psi_x}{V+\phi}\right)|_{x=0} = p_0, \\ (\phi,\psi)(x,0) = (\phi_0,\psi_0)(x) := (v_0(x) - V(0,x), u_0(x) - U(0,x)). \end{cases}$$

Here (V,U) is the smooth approximate solution constructed in section 3.2.

We first choose a positive constant E_0 by virtue of Sobolev's imbedding theorem such that

$$\sup_{x\in R_+}|f(x)|\le\frac{1}{4}v_+ \quad\text{for any}\quad f\in H^1(R_+),\quad \|f\|_1:=\|f\|_{H^1}\le E_0.$$

We look for the solution (ϕ,ψ) of (3.11) in the solution space $X(0,+\infty)$, where

$$\begin{aligned}X(0,T)=\Big\{(\phi,\psi)\,\big|\,&(\phi,\psi)\in C^0\left([0,T];H^1(R_+)\right),\phi_x\in L^2\left([0,T];L^2(R_+)\right)\\ &\psi_x\in L^2\left([0,T];H^1(R_+)\right)\text{and}\sup_{0\le t\le T}\|(\psi,\psi)\|_1\le E_0\Big\}\end{aligned}\tag{3.12}$$

for $0<T\le+\infty$. Note that $V+\phi\ge\frac{3}{4}v_+$ if $(\phi,\psi)\in X(0,T)$.

Similar to the previous papers [9, 6], for the proof of Theorem 3.1 it suffices to show the following theorem.

THEOREM 3.2. *For each fixed constant (v_+,u_+) and $v_-=p^{-1}(p_0)$ $(v_->v_+>0)$ and u_- satisfying* (3.4)*, there exist positive constants ε_0 and C_0 such that if $|(\phi_0,\psi_0)|_1+\delta+|v_0(0)-v_-|\le\varepsilon_0$ $(\delta=|v_+-v_-|+|u_+-u_-|)$, then the problem* (3.11) *has a unique global solution $(\phi,\psi)\in X(0,+\infty)$ which satisfies*

$$\begin{aligned}&\sup_{t\ge0}\|(\phi,\psi)(t)\|_1^2+\int_0^{+\infty}\left(\left\|V_t^{\frac{1}{2}}\phi(\tau)\right\|^2+\|\phi_x(\tau)\|^2+\|\psi_x(\tau)\|_1^2\right)d\tau\\ &\qquad\le C_0\left(\delta^{\frac{1}{6}}+|v_0(0)-v_-|+\|(\phi_0,\psi_0)\|_1^2\right).\end{aligned}\tag{3.13}$$

Along the same lines as in previous papers (see, e.g., [1, 4, 13]), Theorem 3.2 can be shown by the continuation argument combining the local existence with the a priori estimates. For the local existence we define the sequence $\{(\phi^{(n)},\psi^{(n)})\}$ by the iteration

$$(\phi^{(0)},\psi^{(0)})(x)=(\phi_0,\psi_0)(x),$$

$$\begin{cases}\psi_t^{(n+1)}-\mu\left(\dfrac{\psi_x^{(n+1)}}{V+\phi^{(n)}}\right)_x=-(p(V+\phi^{(n)})-p(V))+\mu\left(\dfrac{U_x}{V+\phi^{(n)}}\right)_x,\\ \psi_x^{(n+1)}|_{x=0}=\left\{\dfrac{1}{\mu}(V+\phi^{(n)})(p(V+\phi^{(n)})-p_0)-U_x\right\}\Big|_{x=0},\\ \psi^{(n+1)}(x,0)=\psi_0(x),\end{cases}$$

$$\phi^{(n+1)}(x,t)=\phi_0(x)+\int_0^t\psi_x^{(n+1)}(x,s)ds.$$

Since it is standard to show that $\{(\phi^{(n)},\psi^{(n)})\}$ is the Cauchy sequence in $X(0,t_0)$ for $t_0=t_0(\|(\phi_0,\psi_0)\|_1)>0$ and $\|(\phi_0,\psi_0)\|_1\le E_0$, we omit the details. It now suffices to show the following a priori estimate.

PROPOSITION 3.1 (a priori estimate). *For given constants $(v_\pm,u_\pm)$ in Theorem* 3.2*, suppose that $(\phi,\psi)(x,t)$ is a solution of* (3.11) *in $X(0,T)$ for some $T>0$.*

Then there exist positive constants ε_1 and C_1 independent of T and δ such that if $\sup_{0\le t\le T}\|(\phi,\psi)(t)\|_1+\delta\le\varepsilon_1$, then it holds that

$$\begin{aligned}\sup_{0\le t\le T}\|(\phi,\psi)(t)\|_1^2+\int_0^T\left(\left\|V_t^{\frac12}\phi(\tau)\right\|^2+\|\phi_x(\tau)\|^2+\|\psi_x(\tau)\|_1^2\right)d\tau\\ \le C_1\left(\delta^{\frac16}+|v_0(0)-v_-|+\|(\phi_0,\psi_0)\|_1^2\right).\end{aligned}\tag{3.14}$$

Note that the smallness of $|v_0(0)-v_-|$ in Theorem 3.2 is used in the continuation process.

3.4. A priori estimates. Let $(v_\pm,u_\pm)$ be fixed as in Theorem 3.2 and let $(\phi,\psi)\in X([0,T])$ be a solution of (3.11) for some $T(>0)$ and δ $(0\le\delta<\delta_0)$. Setting $E=\sup_{0\le t\le T}\|(\phi,\psi)(t)\|_1$ and $E(t)=\sup_{0\le\tau\le t}\|(\phi,\psi)(\tau)\|$ we proceed to estimate (ϕ,ψ). Throughout this subsection and later, we write C as generic positive constants independent of T and δ, which may depend on $(v_\pm,u_\pm)$ and v_0.

To prove Proposition 3.1, we need the following estimates at the boundary.

LEMMA 3.3. *Assume $E(t)\le1$. Then it holds for $0\le t\le T$ that*

$$\left|\int_0^t\left(\left(p(V+\phi)-p(V)\right)\psi-\mu\frac{\psi_x\psi}{V+\phi}\right)\Big|_0^{+\infty}d\tau\right|\le\nu\int_0^t\|\psi_x(\tau)\|^2d\tau+C\delta^{\frac43},\tag{3.15}$$

$$\left|\int_0^t\psi(0,\tau)\psi_x(0,\tau)\,d\tau\right|\le\nu\int_0^t\|\psi_x(\tau)\|^2d\tau+C\left(\delta+|v_0-v_-|\right),\tag{3.16}$$

$$\left|\int_0^t\psi_\tau(0,\tau)\psi_x(0,\tau)\,d\tau\right|\le C\left(\delta+|v_0-v_-|\right)\tag{3.17}$$

for any fixed constant $\nu(>0)$.

Proof. Making use of the boundary condition $(3.11)_3$, Lemma 3.2, and Sobolev's inequality, we have

$$\begin{aligned}&\left|\int_0^t\left(\left(p(V+\phi)-p(V)\right)\psi-\mu\tfrac{\psi_x\psi}{V+\phi}\right)\Big|_0^{+\infty}d\tau\right|\\&=\left|\int_0^t\left(p(v(0,\tau))-p(V(0,\tau))-\mu\tfrac{u_x(0,\tau)}{v(0,\tau)}+\mu\tfrac{\mu U_x(0,\tau)}{v(0,\tau)}\right)\psi(0,\tau)\,d\tau\right|\\&=\left|\int_0^t\left(p(v_-)-p(V(0,\tau))+\mu\tfrac{U_x(0,\tau)}{v(0,\tau)}\right)\psi(0,\tau)\,d\tau\right|\\&\le C\int_0^t\left(|v_--V(0,\tau)|+|V_x(0,\tau)|\right)|\psi(0,\tau)|\,d\tau\\&\le C\int_0^t g(\tau)\|\psi(\tau)\|^{\frac12}\|\psi_x(\tau)\|^{\frac12}d\tau\quad\left(\text{where }\ g(t)\triangleq|v_--V(0,t)|+|V_x(0,t)|\right)\\&\le\nu\int_0^t\|\psi_x(\tau)\|^2d\tau+C_\nu\int_0^t g(\tau)^{\frac43}\|\psi(\tau)\|^{\frac23}d\tau\\&\le\nu\int_0^t\|\psi_x(\tau)\|^2d\tau+C_\nu E(t)^{\frac23}\int_0^t g(\tau)^{\frac43}d\tau\\&\le\nu\int_0^t\|\psi_x(\tau)\|^2d\tau+C_\nu\delta^{\frac43},\end{aligned}$$

which shows (3.15). By virtue of (3.15) we derive (3.16) and (3.17) as follows:

$$\begin{aligned}
&\left|\int_0^t \psi(0,\tau)\,\psi_x(0,\tau)\,d\tau\right| \le C\left|\int_0^t \tfrac{\mu\psi(0,\tau)\psi_x(0,\tau)}{v(0,\tau)}d\tau\right| \\
&\le \nu\int_0^t \|\psi_x(\tau)\|^2\,d\tau + C\delta^{\frac{4}{3}} + C\left|\int_0^t (p(V+\phi)-p(V))\,\psi|_0^{+\infty}\,d\tau\right| \\
&\le \nu\int_0^t \|\psi_x(\tau)\|^2\,d\tau + C\delta^{\frac{4}{3}} + C\int_0^t \|\psi(\tau)\|^{\frac{1}{2}}\,\|\psi_x(\tau)\|^{\frac{1}{2}}\,|v(0,\tau)-V(0,\tau)|\,d\tau \\
&\le \nu\int_0^t \|\psi_x(\tau)\|^2\,d\tau + C\delta^{\frac{4}{3}} + CE\int_0^t (|v(0,\tau)-v_-| + |v_- - V(0,\tau)|)\,d\tau \\
&\le \nu\int_0^t \|\psi_x(\tau)\|^2\,d\tau + C\delta^{\frac{4}{3}} + C(|v_0 - v_-| + \delta) \\
&\le \nu\left(\int_0^t \|\psi_x(\tau)\|^2\,d\tau + C(\delta + |v_0 - v_-|)\right)
\end{aligned}$$

and

$$\begin{aligned}
&\left|\int_0^t \psi_\tau(0,\tau)\,\psi_x(0,\tau)\,d\tau\right| \\
&= \left|\int_0^t \psi_\tau(0,\tau)\,(u_x(0,\tau)-U_x(0,\tau))\,d\tau\right| \\
&= \left|\int_0^t \psi_\tau(0,\tau)\,a(\tau)\,d\tau\right| \quad \left(\text{where } a(t) \triangleq \tfrac{1}{\mu}(p(v(0,t))-p(v_-))\,v(0,t) - U_x(0,t)\right) \\
&= \left|a(t)\,\psi(0,t) - a(0)\,\psi(0,0) - \int_0^t \psi(0,\tau)\,a'(\tau)\,d\tau\right| \\
&\le C\left(|a(t)| + |a(0)| + \int_0^t |a'(\tau)|\,d\tau\right)E \\
&\le C(\delta + |v_0(0) - v_-|).
\end{aligned}$$

Thus we complete the proof. □

Using Lemmas 3.2 and 3.3, we can establish the following three lemmas using the same technique as in [11].

LEMMA 3.4. *It follows that for* $0 \le t \le T$,

$$\begin{aligned}
&\|(\phi,\psi)(t)\|^2 + \int_0^t \left(\left\|V_\tau^{\frac{1}{2}}\phi(\tau)\right\|^2 + \|\psi_x(\tau)\|^2\right)d\tau \\
&\le C\left(\delta^{\frac{1}{6}} + \|(\phi_0,\psi_0)\|^2 + E^{\frac{1}{2}}\int_0^t \|(\phi,\psi)_x(\tau)\|^2\,d\tau\right).
\end{aligned}$$

LEMMA 3.5. *It follows that for* $0 \le t \le T$,

$$\begin{aligned}
&\|\phi_x(t)\|^2 + \int_0^t \|\phi_x(\tau)\|^2\,d\tau \\
&\le C\left(\delta^{\frac{1}{2}} + |v_0 - v_-| + \|\phi_{0x}\|^2 + \|\psi_0\|^2 + \|\psi(t)\|^2\right. \\
&\left. + \int_0^t \left(\|\psi_x(\tau)\|^2 + (E+\delta)\,\|\phi_x(\tau)\|^2 + E\,\|\psi_{xx}(\tau)\|^2\right)d\tau\right).
\end{aligned}$$

LEMMA 3.6. *It follows that for* $0 \le t \le T$,

$$\begin{aligned}&\|\psi_x(t)\|^2 + \int_0^t \|\psi_{xx}(\tau)\|^2 d\tau \\ &\le C\left(\delta^{\frac{1}{2}} + |v_0 - v_-| + \|\psi_{0x}\|^2 + \int_0^t \left(\|(\phi,\psi)_x(\tau)\|^2 + E\|\psi_{xx}(\tau)\|^2\right) d\tau\right).\end{aligned}$$

Combining Lemmas 3.4–3.6 yields

$$\begin{aligned}&\|(\phi,\psi)(t)\|_1^2 + \int_0^t \left(\left\|V_\tau^{\frac{1}{2}}\phi(\tau)\right\|^2 + \|\phi_x(\tau)\|^2 + \|\psi_x(\tau)\|_1^2\right) d\tau \\ &\le C\left(\delta^{\frac{1}{6}} + |v_0 - v_-| + \|(\phi_0,\psi_0)\|_1^2\right. \\ &\left. + \int_0^t \left(E^{\frac{1}{2}}\|\psi_x(\tau)\|^2 + \left(E^{\frac{1}{2}} + \delta\right)\|\phi_x(\tau)\|^2 + E\|\psi_{xx}(\tau)\|^2\right) d\tau\right).\end{aligned}$$

Hence, choosing E and δ suitably small as $E + \delta < \varepsilon_1$, we have the a priori estimate (3.14). Thus Proposition 3.1 is completed.

4. Convergence to viscous shock wave. In this section, we discuss the convergence for the solution of (1.6)–(1.8) toward the front viscous shock wave for $p(v) = v^{-\gamma}$ under the condition $v_- < v_+$. Our discussions are largely due to those by Matsumura and Mei [3].

4.1. Viscous shock wave. The viscous shock wave of system (1.6) for the corresponding Cauchy problem is a smooth solution $(V,U)(\xi)$ $(\xi = x - st)$ satisfying (1.6) and $(V,U)(\pm\infty) = (v_\pm, u_\pm)$, namely,

$$\begin{cases} -sV' - U' = 0, \\ -sU' + p(V)' = \mu\left(\frac{U'}{V}\right)', \\ (V,U)(\pm\infty) = (v_\pm, u_\pm), \end{cases} \tag{4.1}$$

where $' = d/d\xi$, s is the shock speed, and $(v_\pm, u_\pm)$ are the given constant states at $\xi = \pm\infty$, satisfying the Rankine–Hugoniot condition

$$\begin{cases} -s(v_+ - v_-) - (u_+ - u_-) = 0, \\ -s(u_+ - u_-) + (p(v_+) - p(v_-)) = 0 \end{cases} \tag{4.2}$$

and the entropy condition

$$\lambda_1(v_+) < s < \lambda_1(v_-)\,(<0) \quad \text{or} \quad (0<)\lambda_2(v_+) < s < \lambda_2(v_-). \tag{4.3}$$

Integrate (4.1) under the Rankine–Hugoniot condition (4.2), and the problem (4.1) is deduced to

$$\begin{cases} \frac{\mu s V'}{V} = -s^2 V - p(V) - b \triangleq h(V), & V(\pm\infty) = v_\pm, \\ U = -s(V - v_\pm) + u_\pm, \end{cases} \tag{4.4}$$

where $b = -s^2 v_\pm - p(v_\pm)$.

In our present problem, the solution to (1.6)–(1.8) is expected to behave as the front viscous shock wave, i.e., $s > 0$, and hence the entropy condition (4.3) yields

$$v_- < v_+. \tag{4.5}$$

Therefore, we have the following lemma on the existence of the front viscous shock wave (see [3]).

LEMMA 4.1. *For any* (v_+, u_+) *and* $v_- = p^{-1}(p_0)$ *with* $0 < v_- < v_+$, *there exist a unique number* $u_- \in R(u_- > u_+)$ *and* $s = \sqrt{-\frac{p(v_+)-p(v_-)}{v_+-v_-}} > 0$ *satisfying* (4.2), *and a unique front viscous shock profile* $(V,U)(\xi)$ $(\xi = x - st)$ *of* (1.6) *up to shift determined by* (4.4), *which satisfies*

$$|V(\xi) - v_\pm, U(\xi) - u_\pm| = O(1)\,|v_+ - v_-, u_+ - u_-|\, e^{-c_\pm|\xi|} \tag{4.6}$$

as $\xi \to \pm\infty$, *where* $c_\pm = v_\pm \left|p'(v_\pm) + s^2\right| / \mu s > 0$.

4.2. Determination of the shift. We first fix a viscous shock wave $(V,U)(x - st)$ mentioned above. Assume that the initial perturbation $(v_0, u_0)(x)$ is given in a neighborhood of the front viscous shock profile $(V,U)(x-\beta)$, where β is a sufficient large constant so that $(V,U)(x-\beta)$ is away from the boundary. Then a shifted front viscous profile $(V,U)(x - st + \alpha - \beta)$ is determined by the data $(v_0, u_0)(x)$ as in [3].

Denote $(V,U) = (V,U)(x - st + \alpha - \beta)$. From $(1.6)_2$ and $(4.1)_2$ we have

$$(u-U)_t = -\left(p(v) - p(V) - \mu\frac{u_x}{v} + \mu\frac{U'}{V}\right)_x. \tag{4.7}$$

Integrating (4.7) over $[0, +\infty)$ with respect to x and using the boundary condition (1.7), we have

$$\begin{aligned} &\frac{d}{dt}\int_0^{+\infty} (u(x,t) - U(x - st + \alpha - \beta))\,dx \\ &= -\left(p(v) - p(V) - \mu\frac{u_x}{v} + \mu\frac{U'}{V}\right)\Big|_0^{+\infty} \\ &= \left(p(v_-) - p(V) + \mu\frac{U'}{V}\right)\Big|_{x=0}. \end{aligned} \tag{4.8}$$

Integrating (4.8) again with respect to t, we get

$$\begin{aligned} &\int_0^{+\infty} (u(x,t) - U(x - st + \alpha - \beta))\,dx \\ &= \int_0^{+\infty} (u_0(x) - U(x + \alpha - \beta))\,dx + \int_0^t \left(p(v_-) - p(V) + \tfrac{\mu U'}{V}\right)\Big|_{x=0} d\tau. \end{aligned} \tag{4.9}$$

If we assume that $u(x,t)$ tends to $U(x - xt + \alpha - \beta)$ in L^1 as $t \to +\infty$, then the right-hand side of (4.9) must go to zero as $t \to +\infty$. Hence, if we set

$$\begin{aligned} I(\alpha) \triangleq &\int_0^{+\infty} (u_0(x) - U(x + \alpha - \beta))\,dx \\ &+ \int_0^{+\infty} \left(p(v_-) - p(V) + \frac{\mu U'}{V}\right)\Big|_{x=0} d\tau, \end{aligned} \tag{4.10}$$

then the shift α must be determined by $I(\alpha) = 0$. Differentiating $I(\alpha)$ with respect to α and using (4.4) yields

$$\begin{aligned} I'(\alpha) &= -\int_0^{\infty} U'(x + \alpha - \beta)\,dx + \int_0^{\infty} (-sU'(-s\tau + \alpha - \beta))\,d\tau \\ &= -u_+ + U(\alpha - \beta) + u_- - U(\alpha - \beta) \\ &= u_- - u_+. \end{aligned} \tag{4.11}$$

Hence $I(\alpha) = I(0) + (u_- - u_+)\alpha$. Thus, the shift $\alpha = \alpha(\beta)$ should be determined explicitly by $I(0) + (u_- - u_+)\alpha = I(\alpha) = 0$, that is,

$$\begin{aligned}\alpha &\triangleq \frac{1}{u_- - u_+}\left(\int_0^\infty (u_0(x) - U(x-\beta))\,dx\right.\\ &\quad\left.+\int_0^\infty (p(v_-) - p(V(-st-\beta)))\,dt + \int_0^\infty \frac{\mu U'}{V}(-st-\beta)\,dt\right)\\ &= \frac{1}{u_- - u_+}\left(\int_0^\infty (u_0(x) - U(x-\beta))\,dx\right.\\ &\quad\left.+\int_0^\infty (p(v_-) - p(V(-st-\beta)))\,dt + \mu\ln\frac{v_-}{V(-\beta)}\right).\end{aligned} \tag{4.12}$$

From (4.9), (4.12), (4.1), and Lemma 4.1 we have heuristically

$$\begin{aligned}&\int_0^\infty (u(x,t) - U(x-st+\alpha-\beta))\,dx\\ &= I(\alpha) - \int_t^\infty \left(p(v_-) - p(V) + \frac{\mu U'}{V}\right)\Big|_{x=0} d\tau\\ &= -\int_t^\infty (p(v_-) - p(V(-s\tau+\alpha-\beta)))\,d\tau - \mu\ln\frac{v_-}{V(-st+\alpha-\beta)}\\ &\to 0 \quad \text{as } t\to\infty.\end{aligned} \tag{4.13}$$

Particularly, note that

$$\begin{aligned}&\int_0^\infty (u_0(x) - U(x+\alpha-\beta))\,dx\\ &= -\int_0^\infty \left(p(v_-) - p(V) + \frac{\mu U'}{V}\right)\Big|_{x=0} dt\\ &= -\int_0^\infty (p(v_-) - p(V(-st+\alpha-\beta)))\,dt + \mu\ln\frac{V(\alpha-\beta)}{v_-}.\end{aligned} \tag{4.14}$$

On the other hand, by the similar argument of $(1.1)_1$ and $(4.1)_1$, the following must hold:

$$\int_0^\infty (v_0(x) - V(x+\alpha-\beta))\,dx + \int_0^\infty (U(-st+\alpha-\beta) - u(0,t))\,dt = 0. \tag{4.15}$$

However, as stated in Matsumura and Mei [3], we expect that $u(0,t)$ is automatically controlled by the effect of boundary so that (4.15) holds with the same shift α defined by (4.12). This situation is really possible because $u(0,t)$ is not specified.

4.3. Main result. Suppose that for some $\beta > 0$,

$$(v_0(x) - V(x-\beta), u_0(x) - U(x-\beta)) \in H^1(R_+) \cap L^1(R_+), \tag{4.16}$$

$$(\Phi_0, \Psi_0)(x) = -\int_x^\infty (v_0(y) - V(y-\beta), u_0(y) - U(y-\beta))\,dy \in L^2(R_+), \tag{4.17}$$

then we have an asymptotic property of the constant shift α as follows.

LEMMA 4.2. *If* (4.16) *and* (4.17) *hold, then* $(\Phi_0, \Psi_0) \in H^2(R_+)$ *and the shift* α *defined by* (4.12) *satisfies that* $\alpha \to 0$ *as* $\|(\Phi_0, \Psi_0)\|_2 \to 0$ *and* $\beta \to +\infty$.

The proof of Lemma 4.2 is easily shown by (4.12).

We now state our second main theorem.

THEOREM 4.1. *For any given* (v_+, u_+), $v_- = p^{-1}(p_0) > 0$ *with* $v_+ > v_-$, *and* u_- *determined in Lemma* 4.1, *suppose the assumptions* (4.16)–(4.17) *and*

$$(\gamma - 1)^2 (v_+ - v_-) < 2\gamma v_-.$$

Then there exists a positive constant ε_2 *such that if* $\|(\Phi_0, \Psi_0)\|_2 + |v_0 - v_-| + \beta^{-1} < \varepsilon_2$, *then* (1.6)–(1.8) *has a unique global solution* $(v, u)(x, t)$ *satisfying*

$$v(x,t) - V(x - st + \alpha - \beta) \in C^0\left([0,\infty); H^1(R_+)\right) \cap L^2\left([0,\infty); H^1(R_+)\right),$$

$$u(x,t) - U(x - st + \alpha - \beta) \in C^0\left([0,\infty); H^1(R_+)\right) \cap L^2\left([0,\infty); H^2(R_+)\right)$$

and, moreover,

$$\sup_{x \in R_+} |(v,u)(x,t) - (V,U)(x - st + \alpha - \beta)| \to 0 \quad \text{as } t \to \infty, \tag{4.18}$$

where $\alpha = \alpha(\beta)$ *is determined by* (4.12).

4.4. Reformulation of the original problem. Define the new unknowns by

$$\begin{cases} \phi(x,t) = -\displaystyle\int_x^\infty (v(y,t) - V(y - st + \alpha - \beta))\, dy, \\ \psi(x,t) = -\displaystyle\int_x^\infty (u(y,t) - U(y - st + \alpha - \beta))\, dy. \end{cases} \tag{4.19}$$

Then the original system (1.6) can be reduced to

$$\begin{cases} \phi_t - \psi_x = 0, \\ \psi_t + (p(V + \phi_x) - p(V)) = \mu\left(\frac{U' + \psi_{xx}}{V + \phi_x} - \frac{U'}{V}\right). \end{cases} \tag{4.20}$$

The initial condition (1.8) is transformed to

$$\begin{aligned} \phi(x,0) &= -\int_x^\infty (v_0(y) - V(y + \alpha - \beta))\, dy \\ &= \Phi_0(x) + \int_x^\infty (V(y + \alpha - \beta) - V(y - \beta))\, dy \\ &= \Phi_0(x) + \int_x^\infty \int_0^\alpha V'(y + \theta - \beta)\, d\theta dy \\ &= \Phi_0(x) + \int_0^\alpha (v_+ - V(x + \theta - \beta))\, d\theta \triangleq \phi_0(x), \end{aligned} \tag{4.21}$$

$$\begin{aligned} \psi(x,0) &= -\int_x^\infty (u_0(y) - U(y + \alpha - \beta))\, dy \\ &= \Psi_0(x) + \int_x^\infty (U(y + \alpha - \beta) - U(y - \beta))\, dy \\ &= \Psi_0(x) + \int_0^\alpha (u_+ - U(x + \theta - \beta))\, d\theta \triangleq \psi_0(x). \end{aligned} \tag{4.22}$$

Then, by the same proof as in [3], we have the following fact for the initial perturbations (4.21) and (4.22) for ϕ and ψ.

LEMMA 4.3. *Under the conditions* (4.16) *and* (4.17)*, the initial perturbation* $(\phi_0, \psi_0) \in H^2(R_+)$ *and satisfies*

$$\|(\phi_0, \psi_0)\|_2 \to 0 \quad \text{as} \quad \|(\Phi_0, \Psi_0)\|_2 \to 0 \text{ and } \beta \to +\infty. \tag{4.23}$$

Concerning the boundary data, from (4.19) and (4.13), it must hold that

$$\begin{aligned} \psi(0,t) &= -\int_0^\infty (u(y,t) - U(y - st + \alpha - \beta))\, dy \\ &= \int_t^\infty (p(v_-) - p(V(-s\tau + \alpha - \beta)))\, d\tau + \mu \ln \tfrac{v_-}{V(-st+\alpha-\beta)} \\ &\triangleq A(t). \end{aligned} \tag{4.24}$$

Thus, by (4.20)–(4.24), we reformulate our problem to

$$\begin{cases} \phi_t - \psi_x = 0, \\ \psi_t - f(V)\phi_x - \frac{\mu}{V}\psi_{xx} = F, \\ (\phi, \psi)(x,0) = (\phi_0, \psi_0)(x) \in H^2(R_+), \ x \geq 0, \\ \psi(0,t) = A(t), \qquad t \geq 0, \end{cases} \tag{4.25}$$

where

$$f(V) = -p'(V) + \frac{\mu s V_x}{V^2} = \frac{h(V) - p'(V)V}{V} \equiv \frac{K(V)}{V}, \tag{4.26}$$

$$F = -(p(V + \phi_x) - p(V) - p'(V)\phi_x) - (\mu\psi_{xx} + h(V)\phi_x)\left(\frac{1}{V + \phi_x} - \frac{1}{V}\right). \tag{4.27}$$

We define the solution space of (4.25) on $I \subset [0, \infty)$ by

$$\begin{aligned} G(I) = \Big\{ &(\phi, \psi) \,\big|\, (\phi, \psi) \in C^0\left(I; H^2(R_+)\right), \ \phi_x \in L^2\left(I; H^1(R_+)\right), \\ &\psi_x \in L^2\left(I; H^2(R_+)\right) \quad \text{with} \quad \sup_{[0,T]} \|(\phi, \psi)(t)\|_2 \leq E_0 \Big\} \end{aligned}$$

for some small constant E_0, so that $\sup_{R_+ \times I}(V + \phi_x)(x,t) \geq \frac{1}{2}v_-$. We also set

$$N(t) = \sup_{0 \leq \tau \leq t} \|(\phi, \psi)(\tau)\|_2, \qquad N_0 = \|\phi_0\|_2 + \|\psi_0\|_2.$$

To prove Theorem 4.1 it suffices to prove the following theorem.

THEOREM 4.2. *Suppose that the assumptions in Theorem* 4.1 *hold. Then there exist positive constants* ε_3 *and* C *such that if* $N_0 + |v_0 - v_-| + \beta^{-1} \leq \varepsilon_3$*, then the initial-boundary value problem* (4.25) *has a unique global solution* $(\phi, \psi) \in G([0, \infty))$ *satisfying*

$$\begin{aligned} &\|(\phi, \psi)(t)\|_2^2 + \int_0^t \left(\|\phi_x(\tau)\|_1^2 + \|\psi_x(\tau)\|_2^2\right) d\tau \\ &\qquad \leq C\left(|v_0 - v_-| + e^{-c_-\beta} + \|(\phi_0, \psi_0)\|_2^2\right), \end{aligned} \tag{4.28}$$

$$\begin{aligned}&\int_0^t \left(\left| \frac{d}{d\tau} \|\phi_x(\tau)\|^2 \right| + \left| \frac{d}{d\tau} \|\psi_x(\tau)\|^2 \right| \right) d\tau \\ &\quad \le C \left(|v_0 - v_-| + e^{-c_-\beta} + \|(\phi_0, \psi_0)\|_2^2 \right)\end{aligned} \tag{4.29}$$

for all $t \ge 0$. Moreover, it holds that the asymptotic stability

$$\sup_{x \in R_+} |(\phi_x, \psi_x)(x,t)| \to 0 \quad \text{as } t \to \infty. \tag{4.30}$$

Similar to Theorem 3.2, Theorem 4.2 can be shown by the continuation argument combining the local existence result together with the a priori estimates. We omit here the local existence result since it is standard. We now give the following a priori estimates which will be shown in the next subsection.

PROPOSITION 4.1 (a priori estimates). *Suppose all assumptions in Theorem* 4.2. *Let $(\phi, \psi) \in G([0,T])$ be a solution to* (4.25) *for $T > 0$. Then there exist positive constants δ_1 and C independent of T such that if $N(T) < \delta_1$, then (ϕ, ψ) satisfies the a priori estimates* (4.28) *and* (4.29) *for $0 \le t \le T$.*

4.5. A priori estimates. Let $(\phi, \psi) \in G([0,T])$ be a solution to (4.25) which satisfies $N(T) < 1$, $\beta > 1$, and $|\alpha| < 1$. In this subsection, we use the letter C to denote some positive constant independent of T, β, and α.

To prove the a priori estimates, we need the following estimates at the boundary.

LEMMA 4.4. *The following inequalities hold for $0 \le t \le T$:*

$$\left| \int_0^t \phi(0,\tau)\psi(0,\tau)\,d\tau \right| \le Ce^{-c_-\beta},$$
$$\left| \int_0^t \psi(0,\tau)\psi_x(0,\tau)\,d\tau \right| \le Ce^{-c_-\beta},$$

$$\left| \int_0^t \phi_x(0,\tau)\psi_x(0,\tau)\,d\tau \right| \le C\left(|v_0 - v_-| + e^{-c_-\beta} \right),$$
$$\left| \int_0^t \psi_x(0,\tau)\psi_t(0,\tau)\,d\tau \right| \le Ce^{-c_-\beta},$$

and

$$\left| \int_0^t \psi_x(0,\tau)\psi_{xx}(0,\tau)\,d\tau \right| \le C\left(|v_0 - v_-| + e^{-c_-\beta} \right),$$
$$\left| \int_0^t \psi_{xt}(0,\tau)\psi_{xx}(0,\tau)\,d\tau \right| \le C\left(|v_0 - v_-| + e^{-c_-\beta} \right).$$

Proof. By Sobolev's inequality, we have

$$\begin{cases} |\phi(0,t)| \le \sup\limits_{x \in R_+} |\phi(x,t)| \le CN(T) \le C, \\ |\psi_x(0,t)| \le \sup\limits_{x \in R_+} |\phi_x(x,t)| \le CN(T) \le C. \end{cases} \tag{4.31}$$

Using Lemma 4.1 gives

$$\begin{aligned}|V(-st + \alpha - \beta) - v_-| &\le Ce^{-c_-|-st+\alpha-\beta|} \\ &= Ce^{-c_-(\beta-\alpha)}e^{-c_-st} \le Ce^{-c_-\beta}e^{-c_-st}.\end{aligned}$$

Thus we have

$$\begin{aligned}|\psi(0,t)| = |A(t)| &\le C\int_t^\infty |V(-s\tau+\alpha-\beta)-v_-|\,d\tau + C\,|V(-st+\alpha-\beta)-v_-| \\ &\le Ce^{-c_-\beta}e^{-c_-st}.\end{aligned}$$

Similarly, we can conclude from (4.4) and Lemma 4.1 that

$$\left|\frac{d^kA(t)}{dt^k}\right| \le Ce^{-c_-\beta}e^{-c_-st}, \qquad k=0,1,2,3. \tag{4.32}$$

From (4.31) and (4.32), it follows that

$$\left|\int_0^t \phi(0,\tau)\,\psi(0,\tau)\,d\tau\right| \le \int_0^t |\phi(0,\tau)|\,|A(\tau)|\,d\tau \le Ce^{-c_-\beta},$$

$$\left|\int_0^t \psi(0,\tau)\,\psi_x(0,\tau)\,d\tau\right| \le \int_0^t |\psi_x(0,\tau)|\,|A(\tau)|\,d\tau \le Ce^{-c_-\beta},$$

$$\left|\int_0^t \psi_x(0,\tau)\,\psi_t(0,\tau)\,d\tau\right| \le \int_0^t |\psi_x(0,\tau)|\,|A'(\tau)|\,d\tau \le Ce^{-c_-\beta}.$$

In light of (4.31), (4.25), (4.32), and Lemma 2.2, we obtain

$$\begin{aligned}\left|\int_0^t \phi_x(0,\tau)\,\psi_x(0,\tau)\,d\tau\right| &\le \int_0^t |v(0,\tau)-V(-s\tau+\alpha-\beta)|\,|\psi_x(0,\tau)|\,d\tau \\ &\le C\int_0^t |v(0,\tau)-V(-s\tau+\alpha-\beta)|\,d\tau \\ &\le C\int_0^t \left(|v(0,\tau)-v_-| + |V(-s\tau+\alpha-\beta)-v_-|\right)d\tau \\ &\le C\left(|v_0-v_-| + e^{-c_-\beta}\right).\end{aligned}$$

By (4.19) and the boundary condition (1.7)

$$\begin{aligned}\psi_{xx}(0,t) &= u_x(0,t) - U'(-st+\alpha-\beta) \\ &= -\tfrac{1}{\mu}v(0,t)\left(p(v(0,t)) - p(v_-)\right) - U'(-st+\alpha-\beta).\end{aligned} \tag{4.33}$$

Hence (4.31) and Lemma 2.2 yield

$$\begin{aligned}\left|\int_0^t \psi_x(0,\tau)\,\psi_{xx}(0,\tau)\,d\tau\right| &\le C\left(\int_0^t \tfrac{1}{\mu}\,|v(0,\tau)|\,|p(v(0,\tau)) - p(v_-)|\,d\tau \right. \\ &\qquad \left. + \int_0^t |U'(-s\tau+\alpha-\beta)|\,d\tau\right) \\ &\le C\left(|v_0-v_-| + e^{-c_-\beta}\right).\end{aligned}$$

Finally, we estimate $|\int_0^t \psi_{xt}(0,\tau)\,\psi_{xx}(0,\tau)\,d\tau|$. Making use of (4.33), (4.31),

Lemma 2.2, (4.4), Lemma 4.1, and the integration of parts, we have

$$\begin{aligned}
&\left|\int_0^t \psi_{xt}(0,\tau)\,\psi_{xx}(0,\tau)\,d\tau\right| \\
=\ &\left|\psi_x(0,t)\,\psi_{xx}(0,t)-\psi_x(0,0)\,\psi_{xx}(0,0)-\int_0^t \psi_x(0,\tau)\,\psi_{xxt}(0,\tau)\,d\tau\right| \\
\le\ &|\psi_x(0,t)\,\psi_{xx}(0,t)|+|\psi_x(0,0)\,\psi_{xx}(0,0)|+\int_0^t |\psi_x(0,\tau)|\,|\psi_{xxt}(0,\tau)|\,d\tau \\
\le\ &C\left(|v_0-v_-|+e^{-c_-\beta}\right)+C\int_0^t \Big(\tfrac{1}{\mu}\,|v_t(0,\tau)|\,|p(v(0,\tau))-p(v_-)| \\
&+\tfrac{1}{\mu}\,|v(0,\tau)|\,|p'(v(0,\tau))\,v_t(0,\tau)|+s\,|U''(-s\tau+\alpha-\beta)|\Big)\,d\tau \\
\le\ &C\left(|v_0-v_-|+e^{-c_-\beta}\right). \qquad \square
\end{aligned}$$

Applying Lemma 4.4, we get the desired a priori estimates (4.28) and (4.29). The proof is the same as that of Proposition 3.4 in [3] except for the boundary values mentioned above, so we omit the details.

REFERENCES

[1] S. Kawashima and A. Matsumura, *Asymptotic stability of travelling wave solutions of systems for one-dimensional gas motion*, Comm. Math. Phys., 101 (1985), pp. 97–127.

[2] A. Matsumura, *Inflow and outflow problems in the half space for a one-dimensional isentropic model system of compressible viscous gas*, in Proceedings of IMS Conference of Differential Equations from Mechanics, Hong Kong, to appear.

[3] A. Matsumura and M. Mei, *Convergence to traveling fronts of solutions of the p-system with viscosity in the presence of a boundary*, Arch. Ration. Mech. Anal., 146 (1999), pp. 1–22.

[4] A. Matsumura and T. Nishida, *The initial value problem for the equations of motion of viscous and heat-conductive gases*, J. Math. Kyoto Univ., 20 (1980), pp. 67–104.

[5] A. Matsumura and K. Nishihara, *On the stability of travelling wave solutions of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 2 (1986), pp. 17–25.

[6] A. Matsumura and K. Nishihara, *Asymptotics toward the rarefaction waves of the solutions of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 3 (1986), pp. 1–13.

[7] A. Matsumura and K. Nishihara, *Global stability of the rarefaction wave of a one-dimensional model system for compressible viscous gas*, Comm. Math. Phys., 144 (1992), pp. 325–335.

[8] A. Matsumura and K. Nishihara, *Global asymptotics toward the rarefaction wave for solutions of viscous p-system with boundary effect*, Quart. Appl. Math., 58 (2000), pp. 69–83.

[9] A. Matsumura and K. Nishihara, *Large-time behaviors of solutions to an inflow problem in the half space for a one-dimensional system of compressible viscous gas*, Comm. Math. Phys., 222 (2001), pp. 449–474.

[10] K. Nishihara, *Asymptotic behaviors of solutions to viscous conservation laws via L^2-energy method*, Adv. Math. (China), 30 (2001), pp. 293–321.

[11] T. Pan, H. Liu, and K. Nishihara, *Asymptotic stability of the rarefaction wave of a one-dimensional model system for compressible viscous gas with boundary*, Japan J. Indust. Appl. Math., 16 (1999), pp. 431–441.

[12] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, New York, Berlin, 1994.

[13] A. I. Vol'pert and S. I. Hujaev, *On the Cauchy problem for composite systems of nonlinear differential equations*, Mat. Sb., 16 (1972), pp. 517–544.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 292–314

EFFECTS OF CERTAIN DEGENERACIES IN THE PREDATOR-PREY MODEL*

E. N. DANCER† AND YIHONG DU‡

Abstract. To demonstrate the influence of spatial heterogeneity on the predator-prey model, we study the effects of the partial vanishing of the nonnegative coefficient functions $b(x)$ and $e(x)$, respectively, in the steady-state predator-prey model

$$\begin{array}{l} -d_1(x)\Delta u = \lambda a_1(x)u - b(x)u^2 - c(x)uv, \\ -d_2(x)\Delta v = \mu a_2(x)c - e(x)v^2 + d(x)uv, \end{array} \qquad u|_{\partial\Omega} = v|_{\partial\Omega} = 0,$$

where all other coefficient functions are strictly positive over the bounded domain Ω in R^N. Critical values of the parameter λ are obtained to show that, in each case, the vanishing has little effect on the behavior of the model when λ is below the critical value, while essential changes occur once λ is beyond the critical value.

Key words. global bifurcation, a priori estimates, predator-prey

AMS subject classifications. 35J20, 35J60

PII. S0036141001387598

1. Introduction.

We are mainly concerned with the nonnegative steady-state solutions of the predator-prey model

$$\left\{ \begin{array}{l} u_t - d_1(x)\Delta u = \lambda a_1(x)u - b(x)u^2 - c(x)uv, \\ v_t - d_2(x)\Delta v = \mu a_2(x)v - e(x)v^2 + d(x)uv, \\ u|_{\partial\Omega\times(0,\infty)} = v|_{\partial\Omega\times(0,\infty)} = 0, \end{array} \right. \tag{1.1}$$

where Ω is a bounded smooth domain in R^N ($N \geq 2$), λ, μ are constants, and $d_1, d_2, a_1, a_2, b, c, d, e$ are nonnegative continuous functions on $\overline{\Omega}$. The dependence on the space variable x of these coefficient functions represents the fact that the prey u and predator v interact in a spatially heterogeneous environment. If the environment is spatially homogeneous, then all these coefficient functions reduce to positive constants, and (1.1) is known as, in this special case, the classical Lotka–Volterra predator-prey model with diffusion, which has attracted extensive study (see, e.g., [BB1, Da1, Da2, KL, Li, LP, Pao, Ya] and the references therein). It is interesting to know whether the model behaves differently when the environment is spatially heterogeneous. When all the coefficient functions are strictly positive over Ω, it is easy to see that (1.1) behaves similarly to the classical Lotka–Volterra case. Thus, we will call (1.1) a *classical* predator-prey model when all the coefficient functions are strictly positive over Ω. A limiting case is when some of the coefficient functions in (1.1) vanish partially over Ω, which we henceforth call a degeneracy. Equation (1.1) with a degeneracy will hence be called a *degenerate* predator-prey model. The main purpose of this paper is to show that the dynamical behavior of (1.1) with certain degeneracies may change drastically from the classical model. This fact shows that the influence of certain

*Received by the editors April 9, 2001; accepted for publication (in revised form) March 9, 2002; published electronically October 8, 2002.
http://www.siam.org/journals/sima/34-2/38759.html
†School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia (normd@maths.usyd.edu.au).
‡School of Mathematical and Computer Sciences, University of New England, Armidale, NSW 2351, Australia (ydu@turing.une.edu.au).

spatial heterogeneity may cause significant changes of behavior for the predator-prey model. A study in the same spirit was carried out recently by Du [Du1, Du2] for the competition model, but the difficulties and the techniques required in the predator-prey model here are very different from those in [Du1, Du2]; the new phenomena revealed are also fundamentally different.

We would like to remark that a degenerate model as described above is a natural limiting problem for the classical model when some of its coefficients are very small on part of the underlying domain. This is an extreme opposite case from all coefficients being constant. Hence a better understanding of the degenerate cases helps the understanding of the more natural classical model with variable coefficients, particularly, the transition of its behavior from a homogeneous environment to an extremely heterogeneous environment. Note that while we understand completely when there is at least one coexistence state for the classical model (see [BB1] or [Da2]), we do not know much about the multiplicity of the coexistence states (except for space dimension 1, where uniqueness is known to hold) and their spatial behavior. To try to understand these problems one is led naturally to degenerate cases. The cases we choose to study here demonstrate considerable differences from the classical homogeneous problem (see, for example, our discussions after Theorem 2.6) and reveal interesting spatial behavior of the coexistence states (see Theorems 2.7 and 3.4). However, they do not seem to lead to multiple coexistence states. It is our hope that this sort of information can also help with the difficult problem of understanding the dynamics of the parabolic predator-prey system.

Let us now describe our results in more detail. We will analyze the effects on the set of steady-state solutions of (1.1) caused by the partial vanishing of $b(x)$ and $e(x)$, respectively. Let us recall that these two functions describe the introspecific pressures of u and v, respectively. The partial vanishing of $b(x)$ implies that in the absence of v, the growth of u is governed by a degenerate logistic law or, more precisely, a mixture of the logistic and Malthusian laws over Ω. The implication of the vanishing of $e(x)$ on v is similar.

When $b(x)$ vanishes partially in Ω, we assume that all the other coefficient functions are positive over Ω. For simplicity, we assume further that all these nonvanishing coefficients are positive constants. It can be easily seen that our arguments work as well without this further assumption. Therefore we lose no generality by doing this. Furthermore, through some simple rescalings of u, v and the coefficients, we see that for the steady-state solutions, we need only consider the following further simplified system:

$$\begin{cases} -\Delta u = \lambda u - b(x)u^2 - cuv, \\ -\Delta v = \mu v - v^2 + duv, \\ u|_{\partial\Omega} = v|_{\partial\Omega} = 0, \end{cases} \tag{1.2}$$

where c, d are positive constants. We assume that $b(x) \equiv 0$ on the closure of some smooth domain Ω_0 satisfying $\overline{\Omega}_0 \subset \Omega$ and $b(x) > 0$ over $\overline{\Omega} \setminus \overline{\Omega}_0$. We will show that there exists a critical value $\lambda^* > 0$ such that (1.2) behaves as if $b(x) \equiv 1$ when $\lambda < \lambda^*$, while essential changes occur once $\lambda \geq \lambda^*$ (see Theorems 2.4 and 2.6 for details). We will also discuss the limiting behavior of the system as $\mu \to -\infty$, where the limiting problem is an interesting free boundary problem (see (2.14), Theorem 2.7, and Remark 2.1 for more details).

Similarly, when $e(x)$ in (1.1) vanishes partially in Ω while all other coefficient functions are positive, then without loss of generality we can consider the following

simplified system:

$$\begin{cases} -\Delta u = \lambda u - u^2 - cuv, \\ -\Delta v = \mu v - e(x)v^2 + duv, \\ u|_{\partial\Omega} = v|_{\partial\Omega} = 0, \end{cases} \tag{1.3}$$

where we assume that $e(x)$ satisfies the same conditions as $b(x)$ given above. Again we will show that there exists a critical number $\lambda_* > 0$ such that (1.3) behaves as if $e(x) \equiv 1$ if $\lambda < \lambda_*$, but drastic changes occur once $\lambda \geq \lambda_*$. However, we will show that the changes now are very different in nature from that for (1.2) (see Theorems 3.2 and 3.3 for details).

The rest of this paper is organized as follows. In section 2 we study (1.2), where for both cases $\lambda < \lambda^*$ and $\lambda \geq \lambda^*$, sufficient and necessary conditions are obtained for the existence of positive solutions of (1.2). Our analysis is based on an a priori estimate result (Theorem 2.1) and a global bifurcation method adapted from [BB2]. When the global bifurcation branch of positive solutions is unbounded, its asymptotic behavior is studied (Theorem 2.7). In section 3 we carry out a similar analysis for (1.3), but as will become clear later, the techniques used and phenomena revealed there are quite different from those in section 2. In the appendix, we collect a few known results used in sections 2 and 3.

2. Degeneracy in the prey equation. This section is devoted to the understanding of the effects of the vanishing of $b(x)$ on the system (1.2). We will, as usual, fix c, d and regard λ and μ as varying parameters. We assume that $b(x)$ possesses the properties described in the introduction.

Clearly $v \equiv 0$ satisfies the second equation in (1.2). In this case u satisfies the so-called degenerate logistic equation

$$-\Delta u = \lambda u - b(x)u^2, \quad u|_{\partial\Omega} = 0. \tag{2.1}$$

It is well known (see [Ou, dP, FKLM]) that (2.1) has only the trivial nonnegative solution $u \equiv 0$ when $\lambda \notin (\lambda_1^\Omega, \lambda_1^{\Omega_0})$, while there is a unique positive solution u_λ when λ belongs to this open interval. Here we use λ_1^ω to denote the first Dirichlet eigenvalue of the Laplacian over the domain ω. For later use, we also introduce the notation $\lambda_1^\omega(\phi)$, which denotes the first eigenvalue of the problem

$$-\Delta u + \phi u = \lambda u, \quad u|_{\partial\omega} = 0.$$

Clearly, $\lambda_1^\omega = \lambda_1^\omega(0)$ under these notations.

It is easily seen that $u_\lambda \to 0$ in $L^\infty(\Omega)$ when $\lambda \to \lambda_1^\Omega$. Moreover, by [DH], as $\lambda \to \lambda_1^{\Omega_0}$,

$$\begin{aligned} &u_\lambda \to \infty && \text{uniformly on } \overline{\Omega}_0, \\ &u_\lambda \to U_{\lambda^{\Omega_0}} && \text{locally uniformly on } \overline{\Omega} \setminus \overline{\Omega}_0, \end{aligned}$$

where U_λ denotes the minimal positive solution of the following boundary blow-up problem:

$$-\Delta U = \lambda U - b(x)U^2, \quad x \in \Omega \setminus \overline{\Omega}_0; \qquad U|_{\partial\Omega} = 0, \quad U|_{\partial\Omega_0} = \infty. \tag{2.2}$$

Here $U|_{\partial\Omega_0} = \infty$ means $\lim_{d(x,\partial\Omega_0)\to 0} U(x) = \infty$. By [DH], (2.2) has a minimal and maximal positive solution for each $\lambda \in (-\infty, \infty)$.

To summarize, for each $\lambda \in (\lambda_1^\Omega, \lambda_1^{\Omega_0})$, (1.2) has a unique semitrivial solution of the form $(u, 0)$ with $u > 0$, namely, $(u_\lambda, 0)$; there is no such semitrivial solution for other λ values.

When $u \equiv 0$, then v satisfies the logistic equation

$$-\Delta v = \mu v - v^2, \quad v|_{\partial\Omega} = 0.$$

It is well known that this equation has no positive solution when $\mu \le \lambda_1^\Omega$, and there is a unique positive solution $v = \theta_\mu$ when $\mu > \lambda_1^\Omega$. Thus (1.2) has a unique semitrivial solution $(0, \theta_\mu)$ of the form $(0, v)$ with $v > 0$ if $\mu > \lambda_1^\Omega$, and there is no such semitrivial solution for other μ values.

The obvious solution $(u, v) = (0, 0)$ of (1.2) is called the trivial solution.

To analyze the set of positive solutions for (1.2) we will need the following a priori estimates.

THEOREM 2.1. *Given an arbitrary positive constant M we can find another positive constant C, depending only on M and b, c, d, Ω in* (1.2)*, such that if (u, v) is a positive solution of* (1.2) *with $|\lambda| + |\mu| \le M$, then*

$$\|u\|_\infty + \|v\|_\infty \le C.$$

Here $\|\cdot\|_\infty = \|\cdot\|_{L^\infty(\Omega)}$.

In the proof of Theorem 2.1, and also in later discussions of the paper, we will need the following result.

LEMMA 2.2. *Suppose $\{u_n\} \subset C^2(\overline{\Omega})$ satisfies*

$$-\Delta u_n \le \lambda u_n, \quad u_n|_{\partial\Omega} = 0, \quad u_n \ge 0, \quad \|u_n\|_\infty = 1,$$

where λ is a positive constant. Then there exists $u_\infty \in L^\infty(\Omega) \cap H_0^1(\Omega)$ such that, subject to a subsequence, $u_n \to u_\infty$ weakly in $H_0^1(\Omega)$, strongly in $L^p(\Omega)$ for all $p \ge 1$, and $\|u_\infty\|_\infty = 1$.

Proof. From the assumption, clearly

$$\int_\Omega |\nabla u_n|^2 dx \le \lambda \int_\Omega u_n^2 dx \le \lambda |\Omega|.$$

Hence $\{u_n\}$ is bounded in $H_0^1(\Omega)$. It follows that by passing to a subsequence, $u_n \to u_\infty$ weakly in $H_0^1(\Omega)$ and strongly in $L^2(\Omega)$. As $\|u_n\|_\infty = 1$, $u_n \to u_\infty$ in $L^2(\Omega)$ implies $u_n \to u_\infty$ in $L^p(\Omega)$ for all $p \ge 1$. Clearly $0 \le u_\infty \le 1$.

It remains to show that $\|u_\infty\|_\infty = 1$. Assume by way of contradiction that $\|u_\infty\|_\infty = 1 - \epsilon < 1$. We are going to prove that this implies $\|u_n\|_\infty < 1$ for all large n, contradicting the assumption that $\|u_n\|_\infty = 1$. This would finish the proof.

From the regularity of the operator $(-\Delta)^{-1}$ we know that

$$w := \lim_{n\to\infty} \lambda(-\Delta)^{-1} u_n = \lambda(-\Delta)^{-1} u_\infty$$

belongs to $C^1(\overline{\Omega})$, and $w = 0$ on $\partial\Omega$. By our assumption, we easily see that $0 \le u_n \le \lambda(-\Delta)^{-1} u_n$. Hence there exists $n_0 \ge 1$ such that for all $n > n_0$, $u_n \le w + 1 - (3/4)\epsilon$. We can now choose a small neighborhood U of $\partial\Omega$ in $\overline{\Omega}$ such that, on U, $u_n < 1 - \epsilon/2$ for $n = 1, \ldots, n_0$ and $w < \epsilon/4$. Thus we have $u_n < 1 - \epsilon/2$ on U for all $n \ge 1$.

In the following, we want to show that for any $x_0 \in \Omega \backslash U$, we can find a small open ball B_{x_0} centered at x_0 such that $u_n \le 1 - \epsilon/2$ on B_{x_0} for all large n. As $\Omega \setminus U$ can

be covered by finitely many such balls, this would eventually mean that $u_n \le 1-\epsilon/2$ on $\Omega \setminus U$ for all large n. Therefore, $\|u_n\|_\infty < 1$ for all large n, as required.

Let us now fix such an x_0. Denote $B_r(x_0) = \{x : |x - x_0| < r\}$ and let

$$v_n(r) = \int_{\partial B_r(x_0)} |u_n(y) - u_\infty(y)| dS_y.$$

We have, for some small positive r_0,

$$\int_0^{r_0} v_n(r)dr = \int_{B_{r_0}(x_0)} |u_n(x) - u_\infty(x)|dx \le \|u_n - u_\infty\|_{L^1(\Omega)} \to 0$$

as $n \to \infty$. Hence, $v_n(r) \to 0$ for almost every $r \in (0, r_0)$.

Choose $r \in (0, r_0)$ very small so that $v_n(r) \to 0$ for this r and that the unique solution to

$$-\Delta w = \lambda, \quad x \in B_r(x_0), \quad w|_{\partial B_r(x_0)} = 0$$

satisfies $\|w\|_\infty < \epsilon/4$. Then let w_n be the unique solution to the problem

$$-\Delta w_n = 0, \quad w_n|_{\partial B_r(x_0)} = u_n.$$

We find that $z_n = w + w_n - u_n$ satisfies

$$-\Delta z_n = \lambda + \Delta u_n \ge \lambda - \lambda u_n \ge 0 \quad \forall x \in B_r(x_0), \quad z_n|_{\partial B_r(x_0)} = 0.$$

Hence, by the maximum principle, $z_n \ge 0$ in $B_r(x_0)$, i.e.,

$$u_n \le w + w_n \quad \forall x \in B_r(x_0).$$

Denote, for $x \in B_r(x_0)$,

$$w_\infty(x) = \frac{r^2 - |x - x_0|^2}{N\omega_N r} \int_{\partial B_r(x_0)} \frac{u_\infty(y)}{|x-y|^N} dS_y,$$

where ω_N stands for the volume of the unit ball in R^N. We clearly have

$$w_\infty(x) \le \frac{r^2 - |x - x_0|^2}{N\omega_N r} \int_{\partial B_r(x_0)} \frac{1-\epsilon}{|x-y|^N} dS_y = 1 - \epsilon \quad \forall x \in B_r(x_0).$$

By the Poisson integral formula, for $x \in B_r(x_0)$,

$$w_n(x) = \frac{r^2 - |x - x_0|^2}{N\omega_N r} \int_{\partial B_r(x_0)} \frac{u_n(y)}{|x-y|^N} dS_y.$$

Therefore, we have, for $x \in B_{r/2}(x_0)$,

$$\begin{aligned} |w_n(x) - w_\infty(x)| &\le \frac{r^2 - |x - x_0|^2}{N\omega_N r} \int_{\partial B_r(x_0)} \frac{|u_n(y) - u_\infty(y)|}{|x-y|^N} dS_y \\ &\le \frac{r^2 - |x - x_0|^2}{N\omega_N r} \int_{\partial B_r(x_0)} \frac{|u_n(y) - u_\infty(y)|}{(r/2)^N} dS_y \\ &\le \frac{2^N}{N\omega_N r^{N-1}} v_n(r) \to 0 \end{aligned}$$

as $n \to \infty$. It follows that for all large n,

$$|w_n(x) - w_\infty(x)| < \epsilon/4 \quad \forall x \in B_{r/2}(x_0).$$

Thus,

$$w_n(x) \le w_\infty(x) + \epsilon/4 \le 1 - (3/4)\epsilon \quad \forall x \in B_{r/2}(x_0).$$

We finally obtain

$$u_n \le w + w_n < \epsilon/4 + 1 - (3/4)\epsilon = 1 - \epsilon/2 \quad \forall x \in B_{r/2}(x_0).$$

This is what we wanted and the proof is thus complete. □

Proof of Theorem 2.1. We use an indirect argument. Suppose the conclusion of our theorem is false. Then we can find a constant $M > 0$ and λ_n, μ_n satisfying

$$|\lambda_n| + |\mu_n| \le M$$

and positive solutions (u_n, v_n) of (1.2) with $\lambda = \lambda_n, \mu = \mu_n$ such that

$$\|u_n\|_\infty + \|v_n\|_\infty \to \infty \quad \text{as } n \to \infty.$$

Since

$$-\Delta v_n = \mu_n v_n - v_n^2 + d u_n v_n \le (M + d\|u_n\|_\infty) v_n - v_n^2,$$

an application of Lemma 2.1 of [DM] (recalled in Lemma A.1 in the appendix) shows that

$$v_n \le M + d\|u_n\|_\infty. \tag{2.3}$$

Therefore we must have $\|u_n\|_\infty \to \infty$.

From the equation for u_n we easily see that $-\Delta u_n \le M u_n$. Hence if we define $\hat{u}_n = u_n/\|u_n\|_\infty$, then

$$-\Delta \hat{u}_n \le M \hat{u}_n. \tag{2.4}$$

Applying Lemma 2.2 we find that, subject to a subsequence, $\hat{u}_n$ converges weakly in $H_0^1(\Omega)$ and strongly in $L^p(\Omega)$ for all $p \ge 1$, to some $\hat{u} \in H_0^1(\Omega)$. Moreover, $\hat{u} \not\equiv 0$.

We claim further that $\hat{u} = 0$ a.e. on $\Omega_+ := \Omega \setminus \overline{\Omega}_0$. To see this, we compare u_n with U_M over Ω_+, where U_M satisfies (2.2) with $\lambda = M$. Clearly we have

$$-\Delta u_n \le M u_n - b(x) u_n^2 \quad \text{over } \Omega_+.$$

Moreover, for each fixed n,

$$\overline{\lim}_{d(x,\partial\Omega_+)\to 0}(u_n - U_M) \le 0.$$

Therefore an application of [DH, Lemma 2.1] gives $u_n \le U_M$ in Ω_+. As $\|u_n\|_\infty \to \infty$, we now easily see that $\hat{u} = 0$ a.e. on Ω_+.

From (2.3), we find that $\hat{v}_n := v_n/\|u_n\|_\infty$ gives rise to a bounded sequence in $L^\infty(\Omega)$. Therefore, by passing to a subsequence, we may assume that $\hat{v}_n$ converges weakly in $L^2(\Omega)$ to some $\hat{v}$. Clearly $\hat{v}$ must be nonnegative and L^∞ bounded.

Choose $\phi \in C_c^\infty(\Omega_0)$ and multiply the equation for u_n by $\phi/\|u_n\|_\infty^2$ and then integrate over Ω_0. We obtain

$$(\|u_n\|_\infty)^{-1} \int_{\Omega_0} \nabla \hat{u}_n \cdot \nabla \phi dx = \int_{\Omega_0} \hat{u}_n(\lambda_n/\|u_n\|_\infty - c\hat{v}_n)\phi dx.$$

Letting $n \to \infty$, we deduce

$$\int_{\Omega_0} \hat{u}\hat{v}\phi dx = 0.$$

This implies

$$\hat{u}\hat{v} = 0 \quad \text{a.e. in } \Omega_0. \tag{2.5}$$

To derive a contradiction, we now look at the equation satisfied by v_n and find that v_n satisfies

$$-\Delta v + \psi v = \mu_n v, \quad v|_{\partial\Omega} = 0,$$

where $\psi = v_n - du_n$. It follows that

$$\mu_n = \lambda_1^\Omega(v_n - du_n). \tag{2.6}$$

By the variational characterization of the first eigenvalue, we have

$$\int_\Omega (|\nabla\phi|^2 + (v_n - du_n)\phi^2)dx \geq \mu_n \int_\Omega \phi^2 dx$$

for any $\phi \in H_0^1(\Omega)$. Choosing $\phi = \hat{u}_n/\sqrt{\|u_n\|_\infty}$, we obtain

$$\int_\Omega |\nabla \hat{u}_n|^2 dx/\|u_n\|_\infty + \int_\Omega (\hat{v}_n - d\hat{u}_n)\hat{u}_n^2 dx \geq \mu_n \int_\Omega \hat{u}_n^2 dx/\|u_n\|_\infty.$$

Letting $n \to \infty$ and recalling $\hat{u}_n \to \hat{u}$ in $L^p(\Omega)$ for any $p \geq 1$, we deduce

$$\int_\Omega (\hat{v}\hat{u}^2 - d\hat{u}^3)dx \geq 0.$$

We already know that $\hat{u} = 0$ a.e. in Ω_+ and $\hat{v}\hat{u} = 0$ a.e. in Ω_0. Therefore $\hat{v}\hat{u}^2 = 0$ a.e. in Ω. It follows that

$$\int_\Omega \hat{u}^3 \leq 0.$$

This can happen only if the nonnegative function $\hat{u}$ is identically zero almost everywhere, which contradicts our earlier observation on $\hat{u}$. This completes the proof of Theorem 2.1. □

We are now ready to study the positive solution set of (1.2). We will adapt the bifurcation approach used by Blat and Brown in [BB2] by fixing λ and using μ as the main bifurcation parameter.

Let us observe that if (1.2) has a positive solution (u, v), then from the first equation in (1.2) we obtain

$$\lambda = \lambda_1^\Omega(bu + cv) > \lambda_1^\Omega(0) = \lambda_1^\Omega.$$

Hence we assume

$$\lambda > \lambda_1^{\Omega}$$

from now on.

Our discussion below is divided into two cases:

$$\text{(i) } \lambda_1^{\Omega} < \lambda < \lambda_1^{\Omega_0} \qquad \text{and} \qquad \text{(ii) } \lambda \geq \lambda_1^{\Omega_0}.$$

In the first case, (1.2) has a unique semitrivial solution of the form $(u, 0)$, namely, $(u_\lambda, 0)$. If (u, v) is a positive solution to (1.2), then u satisfies

$$-\Delta u \leq \lambda u - b(x)u^2, \quad u|_{\partial\Omega} = 0.$$

An application of [DM, Lemma 2.1] (see Lemma A.1) yields

$$0 < u \leq u_\lambda \quad \forall x \in \Omega. \tag{2.7}$$

From the equation for v we find

$$-\Delta v > \mu v - v^2, \quad v|_{\partial\Omega} = 0,$$

which implies, by [DM, Lemma 2.1], that

$$v \geq \theta_\mu \quad \forall x \in \Omega, \tag{2.8}$$

where we make the convention that $\theta_\mu \equiv 0$ whenever $\mu \leq \lambda_1^{\Omega}$.

From the equation for v we also obtain

$$\mu = \lambda_1^{\Omega}(v - du).$$

Therefore, by (2.7) and the well-known monotonicity property of $\lambda_1^{\Omega}(\phi)$, we easily deduce

$$\mu > \lambda_1^{\Omega}(-du_\lambda). \tag{2.9}$$

By the equation for u and (2.8), we deduce

$$\lambda = \lambda_1^{\Omega}(bu + cv) > \lambda_1^{\Omega}(cv) \geq \lambda_1^{\Omega}(c\theta_\mu),$$

that is,

$$\lambda > \lambda_1^{\Omega}(c\theta_\mu). \tag{2.10}$$

Summarizing, we have the following result.

THEOREM 2.3. *In the case that $\lambda_1^{\Omega} < \lambda < \lambda_1^{\Omega_0}$, a necessary condition for* (1.2) *to possess a positive solution is that both* (2.9) *and* (2.10) *hold.*

We will see in the following that (2.9) and (2.10) are also sufficient conditions for the existence of positive solutions. Our argument below is very similar to that of [BB2], and hence we will only sketch it here.

In the (μ, u, v)-space $X := R \times C^1(\overline{\Omega}) \times C^1(\overline{\Omega})$, we have two semitrivial solution curves

$$\Gamma_u := \{(\mu, u_\lambda, 0) : \mu \in (-\infty, \infty)\} \quad \text{and} \quad \Gamma_v := \{(\mu, 0, \theta_\mu) : \lambda_1^{\Omega} < \mu < \infty\}.$$

A local bifurcation analysis along Γ_u shows that a smooth curve of positive solutions $\Gamma' = \{(\mu, u, v)\}$ bifurcates from $(\lambda_1^\Omega(-du_\lambda), u_\lambda, 0) \in \Gamma_u$. A global bifurcation consideration, together with an application of the maximum principle, shows that Γ' is contained in a global branch (i.e., connected set) of positive solutions $\Gamma = \{(\mu, u, v)\}$ which is either unbounded or joins the semitrivial curve Γ_v at exactly $(\mu_0, 0, \theta_{\mu_0}) \in \Gamma_v$, where $\mu_0 > \lambda_1^\Omega$ is determined uniquely by

$$\lambda = \lambda_1^\Omega(c\theta_{\mu_0}). \tag{2.11}$$

It follows from (2.10) that $\mu < \mu_0$ whenever $(\mu, u, v) \in \Gamma$. Therefore, we find that $(\mu, u, v) \in \Gamma$ implies

$$\lambda_1^\Omega(-du_\lambda) < \mu < \mu_0. \tag{2.12}$$

From this, and applying Theorem 2.1, we conclude that Γ is bounded in the space $R \times L^\infty(\Omega) \times L^\infty(\Omega)$. By standard L^p theory for elliptic operators, we conclude that Γ is also bounded in X. Hence Γ must join Γ_v. A local bifurcation analysis near $(\mu_0, 0, \theta_{\mu_0})$ shows that near this point, Γ consists of a smooth curve.

To summarize, we have proved the following result.

THEOREM 2.4. *When $\lambda_1^\Omega < \lambda < \lambda_1^{\Omega_0}$, there is a bounded connected set of positive solutions $\Gamma = \{(\mu, u, v)\}$ in the space X which joins the semitrivial solutions branches Γ_u and Γ_v at $(\lambda_1^\Omega(-du_\lambda), u_\lambda, 0)$ and $(\mu_0, 0, \theta_{\mu_0})$, respectively; moreover, near these two points, Γ consists of smooth curves.*

Clearly, (2.12) is equivalent to (2.9) and (2.10) combined. From Theorems 2.3 and 2.4 the following result now follows.

COROLLARY 2.5. *When $\lambda_1^\Omega < \lambda < \lambda_1^{\Omega_0}$, (1.2) has a positive solution if and only if (2.12) holds.*

The statement in Corollary 2.5 can also be proved by the fixed point index method developed in [Da1, Da2].

In conclusion, we find that our results above are very similar with that for the classical case $b(x) \equiv 1$ obtained in [BB1, BB2] and [Da1, Da2].

Let us now consider the second case where $\lambda \geq \lambda_1^{\Omega_0}$. The striking difference with the classical case now is that we no longer have a semitrivial solution of the form $(u, 0)$. However, the semitrivial solution curve Γ_v is unchanged, and the bifurcation analysis of [BB2] along Γ_v can still be adapted. Again, a local bifurcation analysis shows that a smooth curve of positive solutions $\Gamma' = \{(\mu, u, v)\}$ bifurcates from $(\mu_0, 0, \theta_{\mu_0}) \in \Gamma_v$, where μ_0 is determined by (2.11). As before, a global bifurcation analysis, together with an application of the maximum principle, shows that Γ' is contained in a global branch of positive solutions Γ which is either unbounded in X or joins a semitrivial solution of the form $(u, 0)$. But we already know that there is no semitrivial solution of the form $(u, 0)$. Therefore, Γ must be unbounded.

One easily sees that the arguments leading to (2.10) still work for our present situation. Hence $\mu < \mu_0$ whenever (1.2) has a positive solution. We now apply Theorem 2.1 and conclude that

$$\{\mu : (\mu, u, v) \in \Gamma\} = (-\infty, \mu_0). \tag{2.13}$$

Summarizing the above discussion, we obtain the following result.

THEOREM 2.6. *When $\lambda \geq \lambda_1^{\Omega_0}$, (1.2) has a positive solution if and only if $\mu < \mu_0$. Moreover, there is an unbounded connected set of positive solutions $\Gamma = \{(\mu, u, v)\}$ in X which joins the semitrivial solution branch Γ_v at $(\mu_0, 0, \theta_{\mu_0})$ and satisfies (2.13).*

The fact that (1.2) has a positive solution for arbitrarily large negative μ is strikingly different from the classical case. Biologically, this implies that the prey species can support a predator species of arbitrarily negative growth rate. This is due to the fact that the population of the prey would blow up in the region Ω_0 in the absence of the predator, and hence one might think of Ω_0 as a region where food is abundant for the predator. On the other hand, our above result indicates that the blow-up of the prey population can be avoided by introducing a predator with rather arbitrary growth rate.

It is natural to consider the asymptotic behavior of the positive solutions of (1.2) as $\mu \to -\infty$. For this purpose, we consider a decreasing sequence of negative numbers μ_n which converges to $-\infty$, and we let (u_n, v_n) be an arbitrary positive solution of (1.2) with $\mu = \mu_n$. We show that the following result holds.

THEOREM 2.7. *Let (μ_n, u_n, v_n) be as above. Then the following conclusions are true.*

(i) $\lim_{n\to\infty} \|u_n\|_\infty/|\mu_n| = 1/d$, $\lim_{n\to\infty} \|v_n\|_\infty/|\mu_n| = 0$.

(ii) $u_n/|\mu_n| \to 0$ *and* $v_n \to 0$ *uniformly on any compact subset of* $\overline{\Omega} \setminus \overline{\Omega}_0$.

(iii) $\underline{\lim}_{n\to\infty}\|u_n\|_{L^1(\Omega)}/|\mu_n| > 0$, $\overline{\lim}_{n\to\infty}\|v_n\|_{L^1(\Omega)} < \infty$, *and when* $\lambda > \lambda_1^{\Omega_0}$, $\underline{\lim}_{n\to\infty}\|v_n\|_{L^1(\Omega)} > 0$.

(iv) *If for some* $q > 1$, $\{\|v_n\|_{L^q(\Omega)}\}$ *is bounded, then subject to a subsequence,* $u_n/\|u_n\|_\infty \to \hat{u}$ *weakly in* $H_0^1(\Omega)$, $v_n \to (\lambda/c)\chi_{\{\hat{u}=1\}}$ *weakly in* $L^q(\Omega)$, *where* $\hat{u} = 0$ *a.e. in* $\Omega \setminus \Omega_0$ *and* $\hat{u}|_{\Omega_0}$ *is a positive weak solution* (*with* L^∞ *norm* 1) *of*

$$-\Delta u = \lambda\chi_{\{u<1\}}u, \quad u|_{\partial\Omega_0} = 0. \tag{2.14}$$

Proof. From the equation for v_n we obtain

$$\mu_n > \lambda_1^\Omega - d\|u_n\|_\infty, \tag{2.15}$$

for otherwise,

$$-\Delta v_n \le \lambda_1^\Omega v_n - v_n^2,$$

which gives

$$\int_\Omega |\nabla v_n|^2 dx < \lambda_1^\Omega \int_\Omega v_n^2 dx,$$

contradicting the variational characterization of λ_1^Ω.

From (2.15) we see immediately that $\|u_n\|_\infty \to \infty$ as $n \to \infty$. We can now use (2.4) with $M = \lambda$ and argue as in the proof of Theorem 2.1 to conclude that subject to a subsequence, $\hat{u}_n = u_n/\|u_n\|_\infty \to \hat{u}$ weakly in $H_0^1(\Omega)$ and strongly in $L^p(\Omega)$ for any $p \ge 1$. Moreover, $\hat{u} = 0$ a.e. in Ω_+ and $\hat{u}$ has L^∞ norm 1 over Ω. Furthermore, using standard interior estimates in Ω_+ and boundary estimates near $\partial\Omega$ for the equation satisfied by $\hat{u}_n$, one easily sees that $\hat{u}_n \to 0$ in $C^2(\omega)$ for any compact subset ω of $\overline{\Omega} \setminus \overline{\Omega}_0$. We show next that

$$\lim_{n\to\infty} \mu_n/\|u_n\|_\infty = -d. \tag{2.16}$$

Since

$$-\Delta v_n \le (\mu_n + d\|u_n\|_\infty)v_n - v_n^2,$$

an application of Lemma A.1 shows that

$$v_n \le \mu_n + d\|u_n\|_\infty.$$

It follows that

$$0 \le v_n/\|u_n\|_\infty \le d + \frac{\mu_n}{\|u_n\|_\infty} \le d. \tag{2.17}$$

By passing to a subsequence, we may assume that $\hat{v}_n := v_n/\|u_n\|_\infty$ converges weakly in $L^2(\Omega)$ to $\hat{v}$. The arguments in the proof of Theorem 2.1 which lead to (2.5) work in the same way for our present situation, and hence we still have (2.5), i.e.,

$$\hat{u}\hat{v} = 0 \quad \text{a.e. in } \Omega_0.$$

The identity (2.6) also remains valid and hence

$$\int_\Omega [|\nabla\phi|^2 + (v_n - du_n)\phi^2]dx \ge \mu_n \int_\Omega \phi^2 dx \quad \forall \phi \in H_0^1(\Omega).$$

Dividing the above inequality by $\|u_n\|_\infty$ we obtain, after a simple rearrangement of terms,

$$\int_\Omega \left(\frac{\mu_n}{\|u_n\|_\infty} + d\hat{u}_n - \hat{v}_n\right)\phi^2 dx \le \int_\Omega |\nabla\phi|^2 dx/\|u_n\|_\infty.$$

Letting $n \to \infty$ and denoting $\alpha = \overline{\lim}_{n\to\infty}\mu_n/\|u_n\|_\infty$, we obtain

$$\int_\Omega (\alpha + d\hat{u} - \hat{v})\phi^2 dx \le 0.$$

It follows that

$$\alpha \le -d\hat{u} + \hat{v} \quad \text{a.e. in } \Omega.$$

Since $\hat{u} = 0$ a.e. in Ω_+ and $\hat{u}\hat{v} = 0$ a.e. in Ω_0 and $\|\hat{u}\|_\infty = 1$, we deduce

$$\alpha \le -d.$$

Combining this with (2.15) we find that (2.16) is proved. This finishes the proof of the conclusions about u_n in (i) and (ii). Note that (2.16) holding for a subsequence of an arbitrary subsequence implies that it holds for the entire original sequence.

The second part of (i) follows directly from the first conclusion in (i) and (2.17). We now prove the conclusion about v_n in (ii). By using [DH, Lemma 2.1] we see that $u_n \le U_\lambda$, where U_λ is the minimal positive solution of (2.2). For small $\delta > 0$, define

$$D_\delta = \{x \in \Omega : d(x, \Omega_0) > \delta\}.$$

We find that

$$u_n(x) \le \sup_{x\in D_\delta} U_\lambda(x) = M_\delta < \infty \quad \forall x \in D_\delta.$$

It follows that

$$-\Delta v_n \le (\mu_n + cM_\delta)v_n - v_n^2, \quad x \in D_\delta. \tag{2.18}$$

Denote $a_n = |\mu_n + cM_\delta|$ and define

$$W_n = \frac{\beta}{a_n} d(x)^{-4}, \tag{2.19}$$

where $d(x)$ is a smooth function on $\overline{D}_\delta$ satisfying $d(x) = 0$ on $\partial D_\delta \cap \Omega$ and is positive elsewhere (this is possible if δ is small enough due to the smoothness of $\partial\Omega_0$), and $\beta > 0$ is a constant to be determined later. We calculate

$$\begin{aligned}
&\Delta W_n + (\mu_n + cM_\delta)W_n - W_n^2 \\
&\quad = \frac{\beta}{a_n}\Big(20d(x)^{-6}|\nabla d(x)|^2 - 4d(x)^{-5}\Delta d(x)\Big) - \beta d(x)^{-4} - \frac{\beta^2}{a_n^2} d(x)^{-8} \\
&\quad = \frac{\beta d(x)^{-6}}{a_n}\left(20|\nabla d(x)|^2 - 4d(x)\Delta d(x) - a_n d(x)^2 - \frac{\beta}{a_n} d(x)^{-2}\right) \\
&\quad \le \frac{\beta d(x)^{-6}}{a_n}\Big(20|\nabla d(x)|^2 - 4d(x)\Delta d(x) - 2\sqrt{\beta}\Big) < 0 \quad \forall x \in D_\delta
\end{aligned}$$

if β is chosen large enough.

Thus for such choice of β, for all $n \ge 1$,

$$-\Delta W_n \ge (\mu_n + cM_\delta)W_n - W_n^2 \quad \forall x \in D_\delta.$$

As clearly $W_n > v_n$ on ∂D_δ, we can use (2.18) and Lemma A.1 to conclude that $v_n \le W_n$ on D_δ. Since clearly $W_n \to 0$ uniformly on $D_{2\delta}$, the same is true for v_n. This proves the second part of (ii).

We now consider (iii). If $\underline{\lim}_{n\to\infty} \|u_n\|_{L^1(\Omega)}/|\mu_n| = 0$, then, by passing to a subsequence, we may assume $\|u_n\|_{L^1(\Omega)}/|\mu_n| \to 0$. On the other hand, our previous proof of (i) shows that by passing to a further subsequence, $u_n/|\mu_n| = \hat{u}_n(\|u_n\|_\infty/|\mu_n|) \to \hat{u}/d$ in $L^p(\Omega)$ for all $p \ge 1$, and $\|\hat{u}\|_\infty = 1$. In particular, $\|u_n\|_{L^1(\Omega)}/|\mu_n| \to \|\hat{u}\|_{L^1(\Omega)}/d > 0$. This contradiction shows that we must have $\underline{\lim}_{n\to\infty} \|u_n\|_{L^1(\Omega)}/|\mu_n| > 0$.

If $\overline{\lim}_{n\to\infty} \|v_n\|_{L^1(\Omega)} = \infty$, then by passing to a subsequence, we may assume that $\|v_n\|_{L^1(\Omega)} \to \infty$. As before, by passing to a further subsequence, we have $\hat{u}_n = u_n/\|u_n\|_\infty \to \hat{u}$ in $L^p(\Omega)$ for all $p \ge 1$, and $\|\hat{u}\|_\infty = 1$. Let ϕ be an arbitrary nonnegative function in $C^2(\Omega)$ with compact support in Ω. We multiply the equation for u_n by $\phi/\|u_n\|_\infty$ and integrate over Ω to obtain

$$\int_\Omega \hat{u}_n(-\Delta\phi)dx = \int_\Omega [\lambda\hat{u}_n - b(x)u_n\hat{u}_n - c\hat{u}_n v_n]\phi dx.$$

It follows that

$$\int_\Omega \hat{u}_n v_n \phi dx \le (1/c)\int_\Omega [\lambda\hat{u}_n\phi + \hat{u}_n(\Delta\phi)]dx \to (1/c)\int_\Omega (\lambda\hat{u}\phi + \hat{u}\Delta\phi)dx.$$

Therefore, we can find a positive constant $C_1 = C_1(\phi)$ such that

$$\int_\Omega \hat{u}_n v_n \phi dx \le C_1 \quad \forall n \ge 1. \tag{2.20}$$

Multiplying the equation for v_n by $\phi/|\mu_n|$ and integrating over Ω, we obtain

$$\int_\Omega \frac{v_n}{|\mu_n|}(-\Delta\phi)dx = -\int_\Omega v_n\phi dx + \frac{d\|u_n\|_\infty}{|\mu_n|}\int_\Omega \hat{u}_n v_n\phi dx - \int_\Omega \frac{v_n}{|\mu_n|} v_n\phi dx.$$

Using this identity, (i), and (2.20), we easily deduce

$$\overline{\lim}_{n\to\infty}\int_\Omega v_n\phi dx \le \overline{\lim}_{n\to\infty}\int_\Omega \hat{u}_n v_n\phi dx \le C_1.$$

Thus, there exists some positive constant $C_2 = C_2(\phi)$ such that

$$\int_\Omega v_n\phi dx \le C_2 \quad \forall n \ge 1.$$

By (ii) we know that for any given small closed neighborhood N of $\partial\Omega$ in $\overline{\Omega}$, $v_n \to 0$ uniformly on N. If we choose a particular ϕ in the above discussion such that $\phi \equiv 1$ on $\Omega \setminus N$, then we have

$$\int_{\Omega\setminus N} v_n dx \le \int_\Omega v_n\phi dx \le C_2.$$

Therefore,

$$\|v_n\|_{L^1(\Omega)} = \int_\Omega v_n dx = \int_N v_n dx + \int_{\Omega\setminus N} v_n dx \le C_3 \quad \forall n \ge 1$$

for some positive constant C_3. This contradicts our assumption that $\|v_n\|_{L^1(\Omega)} \to \infty$. Hence we must have $\overline{\lim}_{n\to\infty}\|v_n\|_{L^1(\Omega)} < \infty$.

Suppose now $\lambda > \lambda_1^{\Omega_0}$. To see that $\underline{\lim}_{n\to\infty}\|v_n\|_{L^1(\Omega)} > 0$, we argue indirectly. Suppose this is not true. Then by passing to a subsequence, we may assume that $v_n \to 0$ in $L^1(\Omega)$. Note that on Ω_0, it holds that

$$-\Delta\hat{u}_n = (\lambda - cv_n)\hat{u}_n.$$

We have already proved that subject to a subsequence, $\hat{u}_n \to \hat{u}$ weakly in $H_0^1(\Omega)$, strongly in $L^p(\Omega)$ for all $p \ge 1$, and $\hat{u} \equiv 0$ on Ω_+, and that $\hat{u}$ is positive on a set of positive measure in Ω_0. Moreover, for any $\phi \in C_c^2(\Omega_0)$,

$$\left|\int_\Omega v_n\hat{u}_n\phi dx\right| \le \|\phi\|_\infty \int_\Omega v_n dx \to 0.$$

Thus it is easily seen that $\hat{u}|_{\Omega_0}$ is a weak solution to

$$-\Delta u = \lambda u, \quad u|_{\partial\Omega_0} = 0.$$

As $\hat{u}|_{\Omega_0}$ is nonnegative and not identically zero, we must have $\lambda = \lambda_1^{\Omega_0}$, contradicting our assumption. This finishes the proof for all conclusions in (iii).

It remains to prove (iv). Since by assumption $\{v_n\}$ is bounded in $L^q(\Omega)$, there exists $v \in L^q(\Omega)$ such that, subject to a subsequence, $v_n \to v$ weakly in $L^q(\Omega)$. As before, by passing to a further subsequence, we may assume that $\hat{u}_n = u_n/\|u_n\|_\infty \to \hat{u}$ weakly in $H_0^1(\Omega)$ and strongly in $L^p(\Omega)$ for all $p \ge 1$, and, moreover, $\hat{u} = 0$ a.e. in $\Omega \setminus \Omega_0$ and $\|\hat{u}\|_\infty = 1$. Let ϕ be an arbitrary nonnegative function in $C^2(\Omega)$ with compact support in Ω. Multiply the equation for v_n by $\phi/|\mu_n|$, and integrating over Ω, we deduce

$$\int_\Omega \frac{v_n}{|\mu_n|}(-\Delta\phi)dx = -\int_\Omega v_n\phi dx + \frac{d\|u_n\|_\infty}{|\mu_n|}\int_\Omega \hat{u}_n v_n\phi dx - \int_\Omega \frac{v_n}{|\mu_n|}v_n\phi dx.$$

Making use of (i) and letting $n \to \infty$, we obtain

$$0 = -\int_\Omega v\phi dx + \int_\Omega \hat{u}v\phi dx = -\int_\Omega (1-\hat{u})v\phi dx.$$

As ϕ is arbitrary and $1-\hat{u} \geq 0$ and $v \geq 0$, the above identity implies that $(1-\hat{u})v = 0$ a.e. in Ω. Hence $v = 0$ a.e. in the set $\Omega_1 := \{x \in \Omega : \hat{u}(x) < 1\}$. Since $\hat{u} = 0$ a.e. in $\Omega \setminus \Omega_0$, we find that $\Omega \setminus \Omega_0 \subset \Omega_1$.

Multiplying the equation for u_n by an arbitrary $\psi \in C^2(\Omega_0)$ with compact support in Ω_0 and integrating over Ω_0, we deduce

$$\int_{\Omega_0} \nabla u_n \cdot \nabla \psi dx = \int_{\Omega_0} (\lambda u_n - c u_n v_n)\psi dx.$$

Dividing this identity by $\|u_n\|_\infty$ and letting $n \to \infty$, we obtain

$$\int_{\Omega_0} \nabla \hat{u} \cdot \nabla \psi dx = \int_{\Omega_0} (\lambda - cv)\hat{u}\psi dx.$$

This implies that $\hat{u}|_{\Omega_0}$ is a positive weak solution of the problem

$$-\Delta u = (\lambda - cv)u, \quad u|_{\partial\Omega_0} = 0.$$

Since $(\lambda - cv)\hat{u} \in L^q(\Omega)$, standard elliptic regularity theory implies that $\hat{u}|_{\Omega_0} \in W^{2,q}(\Omega_0)$ and $\Delta\hat{u} = 0$ a.e. in the set $\{x \in \Omega_0 : \hat{u}(x) = 1\}$. Hence $v(x) = \lambda/c$ a.e. in the set $\{x \in \Omega_0 : \hat{u}(x) = 1\}$, which agrees with $\{\hat{u} = 1\} := \{x \in \Omega : \hat{u}(x) = 1\}$, because $\hat{u} = 0$ a.e. in $\Omega \setminus \Omega_0$. Thus,

$$v = (\lambda/c)\chi_{\{\hat{u}=1\}}, \quad 1 - cv = \lambda(1 - \chi_{\{\hat{u}=1\}}) = \lambda\chi_{\{\hat{u}<1\}},$$

and $\hat{u}|_{\Omega_0}$ is a positive weak solution to the problem

$$-\Delta u = \lambda\chi_{\{u<1\}}u, \quad u|_{\partial\Omega_0} = 0.$$

This proves (iv) and hence finishes the proof of Theorem 2.7. □

REMARK 2.1.

(i) *As $\int_\Omega u dx$ and $\int_\Omega v dx$ represent the total population of u and v, respectively, the conclusions in parts* (i) *and* (iii) *of Theorem* 2.7 *imply that, as $\mu \to -\infty$, the total population of u blows up at the rate of $|\mu|$, while the total population of v stays bounded. Moreover, when $\lambda > \lambda_1^{\Omega_0}$, the total population of v is bounded from below by a positive constant independent of μ.*

(ii) *Note that when $\lambda > \lambda_1^{\Omega_0}$, it is easily seen from* (2.14) *that $|\{\hat{u} = 1\}| > 0$ in part* (iv) *of Theorem* 2.7. *Hence $\hat{u}$ has a "flat core." Conversely, if $\lambda = \lambda_1^{\Omega_0}$, then the flat core has measure zero and $\hat{u}$ is the first normalized eigenfunction on Ω_0.*

(iii) *If $\{\|v_n\|_{L^q}\}$ is bounded for some $q > 1$, then from* (2.14) *one sees that $\hat{u} \in C^1(\overline{\Omega_0})$. It is easy to show that $v_n \to 0$ uniformly on any compact subset of the set $\{\hat{u} < 1\}$. However, we were unable to determine whether $\{\|v_n\|_{L^q}\}$ is always bounded for some $q > 1$.*

(iv) *Without the assumption that $\{\|v_n\|_{L^q}\}$ is bounded for some $q > 1$, we can still show that $u_n/\|u_n\|_\infty$ has a subsequence converging to some $\hat{u}$ weakly in $H_0^1(\Omega)$ and that $\hat{u}$ is upper semicontinuous (and hence $\{\hat{u} = 1\}$ is closed) by making use of* [LL, *Theorem* 9.3]. *Moreover, $\hat{u} = 0$ a.e. in $\Omega \setminus \Omega_0$, and on Ω_0, $\hat{u}$ solves $-\Delta u = \lambda\chi_{\{u<1\}}u - m$, where m is a nonnegative measure with support in the boundary of $\{\hat{u} = 1\}$; subject to a subsequence, v_n weak * converges in $C(\Omega)^*$ to $(\lambda/c)\chi_{\{\hat{u}=1\}} + m$.*

(v) *In a forthcoming paper* [DD], *we will show that* $m = 0$ *always, so* $\hat{u}|_{\Omega_0}$ *solves* (2.14). *Moreover, we will show that* (2.14) *has a unique positive solution when* $\lambda \geq \lambda_1^{\Omega_0}$, *and it has no positive solution otherwise. This implies that in part* (iv) *above, the entire sequence* $u_n/\|u_n\|_\infty$ *converges to* $\hat{u}$ *weakly in* $H_0^1(\Omega)$. *A similar comment applies to* v_n.

3. Degeneracy in the predator equation. In this section, we study (1.3). Analogous to the case of (1.2), the semitrivial solutions are easily understood. Now we have a unique semitrivial solution of the form $(0, v)$ for every $\mu \in (\lambda_1^\Omega, \lambda_1^{\Omega_0})$, namely, $(0, v_\mu)$, where v_μ is the unique positive solution of

$$-\Delta v = \mu v - e(x)v^2, \quad v|_{\partial\Omega} = 0; \tag{3.1}$$

and for other values of μ, there is no semitrivial solution of this form.

When $\lambda > \lambda_1^\Omega$, $(\theta_\lambda, 0)$ is the unique semitrivial solution of the form $(u, 0)$, and there is no such semitrivial solution for other λ values.

Our later discussion will require an analysis of $\lambda_1^\Omega(v_\mu)$ as a function of μ.

Let us recall the well-known facts that $\mu \to v_\mu$ is a continuous and strictly increasing function from $(\lambda_1^\Omega, \lambda_1^{\Omega_0})$ to $C(\overline{\Omega})$, $v_\mu \to 0$ as $\mu \to \lambda_1^\Omega$, and as $\mu \to \lambda_1^{\Omega_0}$, by [DH], $v_\mu \to \infty$ uniformly on $\overline{\Omega}_0$ and $v_\mu \to V_{\lambda_1^{\Omega_0}}$ locally uniformly on $\overline{\Omega} \setminus \overline{\Omega}_0$, where V_λ denotes the minimal positive solution of (2.2) with $b(x)$ replaced by $e(x)$.

It follows that

$$\mu \to \Lambda(\mu) := \lambda_1^\Omega(cv_\mu)$$

is a strictly increasing continuous function with $\Lambda(\lambda_1^\Omega - 0) = \lambda_1^\Omega$.

By [Du2, Lemma 3.1] (see Lemma A.2 in the appendix of this paper), we find that

$$\lim_{\mu \to \lambda_1^{\Omega_0}} \Lambda(\mu) = \lambda_* := \lambda_1^{\Omega_+}(cV_{\lambda_1^{\Omega_0}}). \tag{3.2}$$

It follows that

$$\lambda_1^\Omega < \Lambda(\mu) < \lambda_* \quad \forall \mu \in (\lambda_1^\Omega, \lambda_1^{\Omega_0}). \tag{3.3}$$

We will also need the following a priori estimates.

THEOREM 3.1. *Given any positive number* ϵ, *we can find a constant* C, *depending only on* ϵ, c, d, e, *and* Ω, *such that if* (u, v) *is a positive solution of* (1.3) *with* λ, μ *satisfying*

$$|\lambda| + |\mu| \leq \epsilon^{-1}, \quad |\mu - \lambda_1^{\Omega_0}| \geq \epsilon,$$

then

$$\|u\|_\infty + \|v\|_\infty \leq C.$$

Proof. Suppose the conclusion is false. Then we can find some $\epsilon > 0$ and λ_n, μ_n satisfying

$$|\lambda_n| + |\mu_n| \leq \epsilon^{-1}, \quad |\mu_n - \lambda_1^{\Omega_0}| \geq \epsilon \tag{3.4}$$

such that (1.3) with $\lambda = \lambda_n$ and $\mu = \mu_n$ has a positive solution (u_n, v_n) which satisfies

$$\|u_n\|_\infty + \|v_n\|_\infty \to \infty \quad \text{as } n \to \infty.$$

From the equation for u_n we find

$$-\Delta u_n \le \epsilon^{-1} u_n - u_n^2.$$

Comparing u_n with $v \equiv \epsilon^{-1}$ and using Lemma A.1 results in the following:

$$u_n \le \epsilon^{-1} \quad \forall n \ge 1 \quad \forall x \in \Omega. \tag{3.5}$$

Therefore we necessarily have $\|v_n\|_\infty \to \infty$. Let $\hat{v}_n = v_n/\|v_n\|_\infty$. We find, by using the equation for v_n, (3.4), and (3.5), that

$$-\Delta \hat{v}_n \le (1+d)\epsilon^{-1}\hat{v}_n, \quad \hat{v}_n|_{\partial\Omega} = 0.$$

That is, $\hat{v}_n$ satisfies (2.4) with $M = (1+d)\epsilon^{-1}$. Hence we can repeat the argument in the proof of Theorem 2.1 to conclude that, subject to a subsequence, $\hat{v}_n \to \hat{v}$ weakly in $H_0^1(\Omega)$ and strongly in $L^p(\Omega)$ for any $p \ge 1$. Moreover, $\hat{v} = 0$ a.e. in Ω_+ and $\hat{v} > 0$ in a set of positive measure in Ω_0. To show the fact that $\hat{v} = 0$ a.e. in Ω_+, we need the following inequality:

$$-\Delta v_n \le (1+d)\epsilon^{-1} v_n - e(x) v_n^2.$$

Recall that by the smoothness of $\partial\Omega_0$, our above conclusions imply that $w_0 := \hat{v}|_{\Omega_0} \in H_0^1(\Omega_0)$.

By (3.4) and the equation for u_n, we find that u_n satisfies an inequality of the form (2.4), namely,

$$-\Delta u_n \le \epsilon^{-1} u_n.$$

Hence, due to (3.5), subject to a subsequence, $u_n \to u^*$ weakly in $H_0^1(\Omega)$ and strongly in $L^p(\Omega)$ for any $p \ge 1$. It is unclear, however, whether $u^* = 0$ a.e. in Ω.

Let $\phi \in C_c^\infty(\Omega_0)$ and multiply the equation for v_n by $\phi/\|v_n\|_\infty$ and integrate over Ω_0, producing

$$\int_{\Omega_0} \nabla \hat{v}_n \cdot \nabla \phi dx = \mu_n \int_{\Omega_0} \hat{v}_n \phi dx + d \int_{\Omega_0} u_n \hat{v}_n \phi dx.$$

Without loss of generality, we may assume that $\mu_n \to \mu^*$ as $n \to \infty$. We now let $n \to \infty$ in the above identity and deduce

$$\int_{\Omega_0} \nabla \hat{v} \cdot \nabla \phi dx = \int_{\Omega_0} (\mu^* + d u^*) \hat{v} \phi dx.$$

This implies that $w_0 := \hat{v}|_{\Omega_0} \in H_0^1(\Omega)$ is a weak solution to

$$-\Delta w = (\mu^* + d u^*) w, \quad w|_{\partial\Omega_0} = 0. \tag{3.6}$$

Since w_0 is nonnegative and is positive on a set of positive measure, and since $u^* \in L^\infty(\Omega)$, it follows from the Harnack inequality that $w_0 > 0$ in Ω_0.

Let $\psi \in C_c^\infty(\Omega)$ and multiply the equation for u_n by ψ and integrate over Ω. We obtain

$$\int_\Omega \nabla u_n \cdot \nabla \psi dx = \int_\Omega (\lambda_n u_n - u_n^2)\psi dx - c\|v_n\|_\infty \int_\Omega u_n \hat{v}_n \psi dx.$$

It follows easily that

$$\int_\Omega u^* \hat{v} \psi dx = 0.$$

Since $\hat{v}|_{\Omega_0} = w_0 > 0$ in Ω_0, we necessarily have $u^* = 0$ a.e. in Ω_0. Thus (3.6) reduces to

$$-\Delta w_0 = \mu^* w_0, \quad w_0|_{\partial\Omega_0} = 0.$$

This implies that $\mu^* = \lambda_1^{\Omega_0}$, contradicting (3.4).

The proof is complete. □

We are now ready to discuss the set of positive solutions of (1.3). Again we will fix λ and regard μ as the main bifurcation parameter. As before, it is easily seen that when (1.3) has a positive solution, then necessarily, $\lambda > \lambda_1^\Omega$. It turns out that the number $\lambda_* = \lambda_1^{\Omega+}(cV_{\lambda_1^{\Omega_0}})$ is critical in characterizing the behavior of (1.3). So we divide our following discussion into two cases:

$$\text{(i) } \lambda_1^\Omega < \lambda < \lambda_*, \qquad \text{(ii) } \lambda \geq \lambda_*.$$

In case (i), due to the properties of the function $\Lambda(\mu)$, we can find a unique $\mu_0 \in (\lambda_1^\Omega, \lambda_1^{\Omega_0})$ such that

$$\Lambda(\mu_0) = \lambda.$$

A local bifurcation analysis then shows that a smooth curve of positive solutions of (1.3), $\Gamma' = \{(\mu, u, v)\}$, bifurcates from the semitrivial solution curve

$$\Gamma_v := \{(\mu, 0, v_\mu) : \lambda_1^\Omega < \mu < \lambda_1^{\Omega_0}\}$$

at exactly $(\mu_0, 0, v_{\mu_0})$. As before, a global bifurcation consideration, together with an application of the maximum principle, shows that Γ' belongs to a global branch of positive solutions of (1.3), to be denoted by Γ henceforth, which is either unbounded or joins the other semitrivial solution branch

$$\Gamma_u := \{(\mu, \theta_\lambda, 0) : \mu \in (-\infty, \infty)\}$$

at exactly the point $(\lambda_1^\Omega(-d\theta_\lambda), \theta_\lambda, 0) \in \Gamma_u$. Moreover, a local bifurcation analysis shows that when the latter alternative occurs, then Γ consists of a smooth curve near $(\lambda_1^\Omega(-d\theta_\lambda), \theta_\lambda, 0)$.

To determine which alternative occurs for Γ, let us now consider the possible range of μ where (1.3) can have a positive solution. If (u, v) is a positive solution of (1.3), then by using [Du2, Theorem 2.2] for the equation satisfied by v, we deduce

$$\lambda_1^\Omega(-du) < \mu < \lambda_1^{\Omega_0}(-du) < \lambda_1^{\Omega_0}.$$

Since $u \leq \theta_\lambda$, we deduce $\lambda_1^\Omega(-du) \geq \lambda_1^\Omega(-d\theta_\lambda)$. Thus a necessary condition for (1.3) to possess a positive solution is

$$\lambda_1^\Omega(-d\theta_\lambda) < \mu < \lambda_1^{\Omega_0}. \tag{3.7}$$

From the equation for u we obtain

$$\lambda = \lambda_1^\Omega(u + cv).$$

An application of Lemma A.1 for the equation of v gives $v \geq v_\mu$, where we define $v_\mu = 0$ for $\mu \leq \lambda_1^\Omega$. Therefore,

$$\lambda = \lambda_1^\Omega(u + cv) > \lambda_1^\Omega(cv_\mu).$$

This implies $\mu < \mu_0$, and hence the necessary condition (3.7) can be improved to

$$\lambda_1^\Omega(-d\theta_\lambda) < \mu < \mu_0. \tag{3.8}$$

We can now apply Theorem 3.1 to conclude that Γ must be bounded. Hence, we have the following result.

THEOREM 3.2. *When $\lambda_1^\Omega < \lambda < \lambda_*$, (1.3) has a positive solution if and only if (3.8) holds. Moreover, there is a bounded connected set of positive solutions of (1.3), $\Gamma = \{(\mu, u, v)\}$, which joins the semitrivial solution branches Γ_u and Γ_v at $(\lambda_1^\Omega(-d\theta_\lambda), \theta_\lambda, 0)$ and $(\mu_0, 0, v_{\mu_0})$, respectively.*

The above result is clearly very similar to that for the classical case $e(x) \equiv 1$ obtained in [BB1, BB2] and [Da1, Da2]. We show in the following that for the case $\lambda \geq \lambda_*$, very different behavior arises for (1.3).

Suppose from now on that $\lambda \geq \lambda_*$. A local and global bifurcation analysis much as before shows that from the semitrivial solution curve Γ_u, a global bifurcation branch $\Gamma = \{(\mu, u, v)\}$ of positive solutions of (1.3) bifurcates from the point $(\lambda_1^\Omega(-d\theta_\lambda), \theta_\lambda, 0)$, and it is either unbounded or joins the semitrivial solution curve Γ_v. In the latter case, we can find a sequence $(\mu_n, u_n, v_n) \in \Gamma$ such that $(\mu_n, u_n, v_n) \to (\mu, 0, v_\mu) \in \Gamma_v$ in the space $R \times C(\overline{\Omega}) \times C(\overline{\Omega})$. Then, from the equation for u_n and (3.3), we obtain

$$\lambda = \lambda_1^\Omega(u_n + cv_n) \to \lambda_1^\Omega(cv_\mu) = \Lambda(\mu) < \lambda_*.$$

This contradicts our assumption that $\lambda \geq \lambda_*$. Therefore Γ must be unbounded.

One easily checks that the analysis leading to (3.7) is still valid for our present situation. Hence, (3.7) is still a necessary condition and by Theorem 3.1, Γ contains points (μ_n, u_n, v_n) such that $\mu_n \to \lambda_1^{\Omega_0}$. Summarizing, we obtain the following result.

THEOREM 3.3. *When $\lambda \geq \lambda_*$, (1.3) has a positive solution if and only if (3.7) holds. Moreover, there is an unbounded connected set of positive solutions of (1.3), $\Gamma = \{(\mu, u, v)\}$, which joins the semitrivial solution curve Γ_u at $(\lambda_1^\Omega(-d\theta_\lambda), \theta_\lambda, 0)$ and remains bounded until μ approaches $\lambda_1^{\Omega_0}$, where it blows up. Γ does not join the other semitrivial solution curve Γ_v.*

To analyze the blow-up behavior of the positive solutions of (1.3) as $\mu \to \lambda_1^{\Omega_0}$, we make the following further assumption on $e(x)$:

$$\lim_{d(x,\Omega_0)\to 0} \frac{e(x)}{d(x,\Omega_0)^\alpha} = c \quad \text{for some positive constants } \alpha \text{ and } c. \tag{3.9}$$

By [DH, Theorem 2.8], we know that (3.9) guarantees that problem (2.2) with $b(x)$ replaced by $e(x)$ has a unique positive solution V_λ.

Suppose μ_n is an increasing sequence of positive numbers converging to $\lambda_1^{\Omega_0}$ as $n \to \infty$, and (u_n, v_n) is an arbitrary positive solution of (1.3) with $\mu = \mu_n$. We have the following result which describes the asymptotic behavior of (u_n, v_n).

THEOREM 3.4. *Suppose (3.9) holds and $\lambda \geq \lambda_*$. Then the following are true.*

(i) $\lim_{n\to\infty} u_n = 0$ *uniformly on any compact subset of* Ω_0.

(ii) $\lim_{n\to\infty} v_n = \infty$ *uniformly on* $\overline{\Omega}_0$.

(iii) *If* $\lambda = \lambda_*$, *then* $\lim_{n\to\infty}(u_n, v_n)|_{\Omega_+} = (0, V_{\lambda_1^{\Omega_0}})$ *in the space* $C(\overline{\Omega}_+)\times C^1_{\rm loc}(\Omega_+\cup\partial\Omega)$, *where* $C^1_{\rm loc}(\Omega_+\cup\partial\Omega) = \cap_D C^1(D)$ *with* D *running through all the closed subsets of* $\Omega_+\cup\partial\Omega$.

(iv) *If* $\lambda > \lambda_*$, *then, subject to a subsequence,* $\lim_{n\to\infty}(u_n, v_n)|_{\Omega_+} = (u^*, v^*)$ *in the space* $C^1_{\rm loc}(\Omega_+\cup\partial\Omega)\times C^1_{\rm loc}(\Omega_+\cup\partial\Omega)$, *where* (u^*, v^*) *is a positive solution of the boundary blow-up problem*

$$\begin{cases} -\Delta u = \lambda u - u^2 - cuv, \\ -\Delta v = \lambda_1^{\Omega_0} v - e(x)v^2 + duv, \\ u|_{\partial\Omega_+} = 0, \quad v|_{\partial\Omega} = 0, \quad v|_{\partial\Omega_0} = \infty. \end{cases} \quad x \in \Omega_+, \tag{3.10}$$

Proof. Denote $w_n = v_{\mu_n}$. Then it follows from

$$-\Delta v_n \geq \mu_n v_n - e(x)v_n^2, \quad v_n|_{\partial\Omega} = 0,$$

and [DH, Lemma 2.1] that $w_n \leq v_n$. By [DH, Theorem 3.6], $w_n \to \infty$ uniformly on $\overline{\Omega}_0$. Hence so is v_n. This proves (ii).

Denote $\eta_n = \min_{x\in\overline{\Omega}_0} v_n(x)$. The above proved conclusion (ii) implies that $\eta_n \to \infty$ as $n \to \infty$. From the equation for u_n we find

$$-\Delta u_n \leq (\lambda - \eta_n)u_n - u_n^2 \quad \forall x \in \Omega_0. \tag{3.11}$$

Let W_n be defined by (2.19) but with $a_n = |\lambda - \eta_n|$ and D_δ replaced by Ω_0. Then the same calculation as in the proof of Theorem 2.7 shows

$$-\Delta W_n \geq (\lambda - \eta_n)W_n - W_n^2 \quad \forall x \in \Omega_0, \tag{3.12}$$

provided that $\beta > 0$ is chosen large enough.

By (3.11), (3.12), and Lemma A.1, we deduce that $u_n \leq W_n$ and (i) follows.

Since $u_n \leq \theta_\lambda$, much as before, it follows from

$$-\Delta u_n \leq \lambda u_n, \quad u_n|_{\partial\Omega} = 0$$

that, subject to a subsequence, $u_n \to u^*$ weakly in $H_0^1(\Omega)$ and strongly in $L^p(\Omega)$ for any $p \geq 1$. By conclusion (i) we see that $u^* = 0$ a.e. in Ω_0. It follows from the smoothness assumption on $\partial\Omega_0$ that

$$u^*|_{\Omega_+} \in H_0^1(\Omega_+). \tag{3.13}$$

To prove (iii), we assume now that $\lambda = \lambda_*$. We first show that $u^* = 0$ a.e. in Ω. Otherwise, due to (i), $u^* > 0$ on a set of positive measure in Ω_0. Recalling that $v_n \geq w_n$, we deduce from the equation of u_n that

$$\lambda = \lambda_1^\Omega(u_n + cv_n) \geq \lambda_1^\Omega(u_n + cw_n).$$

Using the properties of w_n and applying Lemma A.2 in the appendix, we deduce

$$\lambda_1^\Omega(u_n + cw_n) \to \lambda_1^{\Omega_+}(u^* + cV_{\lambda_1^{\Omega_0}}) > \lambda_*.$$

This implies that $\lambda > \lambda_*$, contradicting our assumption that $\lambda = \lambda_*$. This proves $u^* = 0$ a.e. in Ω. Hence $u_n \to 0$ in $L^p(\Omega)$ for any $p \geq 1$.

We claim further that $u_n \to 0$ in $L^\infty(\Omega)$. Indeed, from $-\Delta u_n \leq u_n$ we deduce

$$0 \leq u_n \leq \lambda(-\Delta)^{-1}u_n.$$

By using the regularity of the operator $(-\Delta)^{-1}$ repeatedly, we easily see from the above inequality that $u_n \to 0$ in $L^p(\Omega)$ for any $p \geq 1$ implies that $u_n \to 0$ in $L^\infty(\Omega)$.

Next we show that $v_n \to V_{\lambda_1^{\Omega_0}}$ in $C^1_{\rm loc}(\Omega_+ \cup \partial\Omega)$. To this end, we consider a sequence of enlarging smooth domains Ω_n given by

$$\Omega_n = \{x \in \Omega_0 : d(x, \partial\Omega_0) > \delta_n\},$$

where δ_n is a decreasing sequence of positive numbers converging to 0. We assume that δ_1 is small enough such that for each $n \geq 1$, Ω_n is not empty and $\partial\Omega_n$ is as smooth as $\partial\Omega_0$.

Let

$$e_n(x) = e(x) + d(x, \Omega_n), \quad x \in \Omega.$$

Clearly $e_n(x)$ has the following properties:

(a) $e_n \to e$ in $L^\infty(\Omega)$,
(b) $e_n(x) > 0$ on $\overline{\Omega} \setminus \overline{\Omega}_n$,
(c) $e_n(x) = 0$ on $\overline{\Omega}_n$, and
(d) $e_n(x) \geq e_{n+1}(x)$ for all $x \in \Omega$.

For fixed $\epsilon > 0$, by [DH, Theorem 2.8], for each n there is a unique positive solution Z_n to the problem

$$-\Delta Z = (\mu_n + \epsilon)Z - e_n(x)Z^2 \quad \text{in } \Omega \setminus \overline{\Omega}_n, \quad Z|_{\partial\Omega} = 0, \quad Z|_{\partial\Omega_n} = \infty.$$

By [DH, Lemma 2.1], we find that on Ω_+, Z_n increases with n, and hence $Z^*(x) = \lim_{n\to\infty} Z_n(x)$ is well defined over Ω_+. A simple regularity consideration reveals that Z^* is a solution to

$$-\Delta Z = (\lambda_1^{\Omega_0} + \epsilon)Z - e(x)Z^2, \quad Z|_{\partial\Omega} = 0, \quad Z|_{\partial\Omega_0} = \infty.$$

As $V_{\lambda_1^{\Omega_0}+\epsilon}$ is the unique positive solution to this problem, we must have $Z^* = V_{\lambda_1^{\Omega_0}+\epsilon}$, that is,

$$\lim_{n\to\infty} Z_n(x) = V_{\lambda_1^{\Omega_0}+\epsilon}(x) \quad \forall x \in \Omega_+. \tag{3.14}$$

Since $\|u_n\|_\infty \to 0$ as $n \to \infty$, for all large n, we obtain from the equation for v_n that

$$-\Delta v_n = (\mu_n + du_n)v_n - e(x)v_n^2 \leq (\mu_n + \epsilon)v_n - e_n(x)v_n^2.$$

Apply [DH, Lemma 2.1], we find $v_n \leq Z_n$ on $\Omega_+ \subset \Omega \setminus \overline{\Omega}_n$ for all large n. Using (3.14), we deduce

$$\overline{\lim}_{n\to\infty} v_n(x) \leq V_{\lambda_1^{\Omega_0}+\epsilon}(x) \quad \forall x \in \Omega_+. \tag{3.15}$$

The uniqueness of V_λ implies that V_λ varies continuously with respect to λ in the norm $C^1_{\rm loc}(\Omega_+ \cup \partial\Omega)$, which follows from a simple regularity and compactness consideration. As (3.15) holds for all small $\epsilon > 0$, letting $\epsilon \to 0$, we deduce

$$\overline{\lim}_{n\to\infty} v_n(x) \leq V_{\lambda_1^{\Omega_0}}(x) \quad \forall x \in \Omega_+.$$

On the other hand, we already know that $v_n(x) \geq w_n(x)$ and $w_n(x) \to V_{\lambda_1^{\Omega_0}}(x)$, so it follows that

$$\underline{\lim}_{n\to\infty} v_n(x) \geq V_{\lambda_1^{\Omega_0}}(x) \quad \forall x \in \Omega_+.$$

Therefore we must have

$$\lim_{n\to\infty} v_n(x) = V_{\lambda_1^{\Omega_0}}(x) \quad \forall x \in \Omega_+.$$

By a simple regularity consideration, we find that $v_n \to V_{\lambda_1^{\Omega_0}}$ in $C^1_{\rm loc}(\Omega_+ \cup \partial\Omega)$. This finishes the proof of (iii).

Finally we consider the case $\lambda > \lambda_*$ and prove (iv). Our first observation is that $u^* > 0$ on a set of positive measure in Ω_0. Otherwise, we can follow the argument in the proof of (iii) above to conclude that $u_n \to 0$ in $L^\infty(\Omega)$ and $v_n \to V_{\lambda_1^{\Omega_0}}$ in $C^1_{\rm loc}(\Omega_+ \cup \partial\Omega)$. Hence, by Lemma A.2,

$$\lambda = \lambda_1^{\Omega}(u_n + cv_n) \to \lambda_1^{\Omega_+}(cV_{\lambda_1^{\Omega_0}}) = \lambda_*.$$

This contradicts our assumption that $\lambda > \lambda_*$.

If $\eta > \lambda_1^{\Omega_0} + d\|\theta_\lambda\|_\infty$, then

$$-\Delta v_n = \mu_n - e(x)v_n^2 + du_nv_n \leq \eta v_n - e(x)v_n^2.$$

Therefore we can use [DH, Lemma 2.1] to conclude that $v_n \leq V_\eta$ on Ω_+, where we recall that V_η is the unique positive solution of (2.2) with $\lambda = \eta$ and $b(x) = e(x)$. This implies that $\{v_n(x)\}$ is uniformly bounded on any compact subset of $\Omega_+ \cup \partial\Omega$. Define $D_n = \{x \in \overline{\Omega} : d(x, \Omega_0) > \delta_n\}$, where δ_n is a decreasing sequence of positive numbers converging to 0; we easily see, using interior and boundary estimates on D_{K+1}, that $\{v_n|_{D_k}\}$ is compact in $C^1(D_k)$. Hence, by a standard diagonal process, $\{v_n|_{\Omega_+}\}$ has a subsequence converging to some v^* in $C^1_{\rm loc}(\Omega_+ \cup \partial\Omega)$. Clearly $v^*|_{\partial\Omega} = 0$. Since $w_n(x) \leq v_n(x)$ and $w_n(x) \to V_{\lambda_1^{\Omega_0}}(x)$ on Ω_+, we deduce $v^* \geq V_{\lambda_1^{\Omega_0}}$ on Ω_+. Thus $v^* > 0$ on Ω_+ and $v^*|_{\partial\Omega_0} = \infty$. By passing to the limit along a proper subsequence in the equations for u_n and v_n, it is now easily seen that $(u^*, v^*)|_{\Omega_+}$ satisfies (3.10). By standard interior and boundary regularity, we find that $u^*|_{\Omega_+}$ belongs to $C^1_{\rm loc}(\Omega_+ \cup \partial\Omega)$, and by applying the Harnack inequality on each compact subset of $\Omega_+ \cup \partial\Omega$, we find that $u^* > 0$ in Ω_+. This proves (iv) and hence completes the proof of Theorem 3.4. □

Theorem 3.4 implies that when the growth rate of the predator approaches a certain critical positive value (i.e., $\lambda_1^{\Omega_0}$), then the predator population blows up in the region Ω_0 and the prey population vanishes in Ω_0 due to the strong presence of the predator in that region. Note, however, that the prey population survives in Ω_+. A positive growth rate of the predator implies that it has other food supplies apart from the prey u, which seems reasonable in many practical situations.

Remark 3.1.

(i) *As in the classical case, we suspect that for* (1.2) *and* (1.3), *there is at most one positive solution, which is the global attractor of the positive solutions of the corresponding parabolic system.*

(ii) *Our results are valid if* Ω *is a finite interval. In this case,* Ω_+ *is no longer connected; instead, it becomes the union of two disconnected intervals. Hence* (2.2) *and* (3.10) *split into problems on these two subintervals. The conclusions remain valid on each subinterval.*

(iii) *A direct adaptation of the argument in* [LP] *shows that in the case* Ω *is a finite interval,* (1.2) *and* (1.3) *have at most one positive solution. It follows that the connected set of positive solutions,* Γ*, in Theorems* 2.4, 2.6, 3.2, *and* 3.3 *is a smooth curve and contains all the positive solutions.*

(iv) *All our existence results can be obtained by the fixed point index approach developed in Dancer* [Da1, Da2].

Appendix. For the convenience of the reader, we state precisely [DM, Lemma 2.1] and [Du2, Lemma 3.1] here.

LEMMA A.1 (see [DM, Lemma 2.1]). *Suppose that* Ω *is a bounded domain in* R^N*,* $\alpha(x)$ *and* $\beta(x)$ *are continuous functions on* Ω *with* $\|\alpha\|_{L^\infty(\Omega)} < \infty$*, and* $\beta(x)$ *is nonnegative and not identically zero. Let* $u_1, u_2 \in C^2(\Omega)$ *be positive in* Ω *and satisfy*

$$Lu_1 + \alpha(x)u_1 - \beta(x)g(u_1) \le 0 \le Lu_2 + \alpha(x)u_2 - \beta(x)g(u_2), \quad x \in \Omega,$$

and $\overline{\lim}_{x\to\partial\Omega}(u_2 - u_1) \le 0$*, where* $Lu = \Sigma_{ij}[a_{ij}(x)u_{x_i}]_{x_j}$ *is a uniformly elliptic operator with smooth coefficients* a_{ij}*, and* $g(u)$ *is continuous and such that* $g(u)/u$ *is strictly increasing for* u *in the range* $\inf_\Omega\{u_1, u_2\} < u < \sup_\Omega \|u_1, u_2\}$*. Then* $u_2 \le u_1$ *in* Ω*.*

Note that the positive functions u_1 and u_2 may not be defined on $\partial\Omega$. Therefore this comparison result can be applied to solutions with boundary blow-ups. The existence of such *positive* functions u_1 and u_2 has hidden restrictions on $\alpha(x)$ and $\beta(x)$.

Let $\lambda_1^\omega(\phi)$ be as defined at the beginning of section 2, and let Ω, Ω_0, and Ω_+ be the same as used in this paper. Then the following result holds.

LEMMA A.2 (see [Du2, Lemma 3.1]). *Suppose that* $\phi_n \in C(\overline{\Omega})$ *and satisfies*

(i) $\phi_n \ge -M$ *for some positive constant* M*,* $\phi_n \to \infty$ *uniformly on* $\overline{\Omega}_0$ *as* $n \to \infty$*, and*

(ii) $\phi_n \to \phi$ *in* $L^p(\Omega')$ *for all* $p > 1$ *and* $\Omega' \subset\subset \Omega_+$*, where* $\phi \in C(\Omega_+ \cup \partial\Omega)$*.*

Then $\lambda_1^\Omega(\phi_n) \to \lambda_1^{\Omega_+}(\phi)$*.*

REFERENCES

[BB1] J. BLAT AND K.J. BROWN, *Bifurcation of steady-state solutions in predator-prey and competition systems*, Proc. Roy. Soc. Edinburgh Sect. A, 97 (1984), pp. 21–34.

[BB2] J. BLAT AND K.J. BROWN, *Global bifurcation of positive solutions in some systems of elliptic equations*, SIAM J. Math. Anal., 17 (1986), pp. 1339–1353.

[Da1] E.N. DANCER, *On positive solutions of some pairs of differential equations*, Trans. Amer. Math. Soc., 284 (1984), pp. 729–743.

[Da2] E.N. DANCER, *On positive solutions of some pairs of differential equations*, II, J. Differential Equations, 60 (1985), pp. 236–258.

[DD] E.N. DANCER AND Y. DU, *On a free boundary problem arising from population biology*, Indiana Univ. Math. J., to appear.

[dP] M.A. DEL PINO, *Positive solutions of a semilinear equation on a compact manifold*, Nonlinear Anal., 22 (1994), pp. 1423–1430.

[Du1] Y. DU, *Effects of a degeneracy in the competition model, part* I: *Classical and generalized steady-state solutions*, J. Differential Equations, 181 (2002), pp. 92–132.

[Du2] Y. DU, *Effects of a degeneracy in the competition model, part* II: *Perturbation and dynamical behaviour*, J. Differential Equations, 181 (2002), pp. 133–164.

[DH] Y. DU AND Q. HUANG, *Blow-up solutions for a class of semilinear elliptic and parabolic equations*, SIAM J. Math. Anal., 31 (1999), pp. 1–18.

[DM] Y. DU AND L. MA, *Logistic type equations on* R^N *by a squeezing method involving boundary blow-up solutions*, J. London Math. Soc., 64 (2001), pp. 107–124.

[FKLM] J.M. FRAILE, P. KOCH MEDINA, J. LOPEZ-GOMEZ, AND S. MERINO, *Elliptic eigenvalue problems and unbounded continua of positive solutions of a semilinear elliptic equation*, J. Differential Equations, 127 (1996), pp. 295–319.

[KL] P. Korman and A.W. Leung, *A general monotone scheme for elliptic system with applications to ecological models*, Proc. Roy. Soc. Edinburgh Sect. A, 102 (1986), pp. 315–325.

[Li] L. Li, *Coexistence theorems of steady-states for predator-prey interacting systems*, Trans. Amer. Math. Soc., 305 (1988), pp. 143–166.

[LL] E.H. Lieb and M. Loss, *Analysis*, AMS, Providence, RI, 1997.

[LP] J. Lopez-Gomez and R. Pado, *Existence and uniqueness of coexistence states for the predator-prey model with diffusion: The scalar case*, Differential Integral Equations, 6 (1993), pp. 1025–1031.

[Pao] C.V. Pao, *On nonlinear reaction-diffusion systems*, J. Math. Anal. Appl., 87 (1982), pp. 165–198.

[Ou] T. Ouyang, *On the positive solutions of semilinear equation* $\triangle u + \lambda u - hu^p = 0$ *on the compact manifolds*, Trans. Amer. Math. Soc., 331 (1992), pp. 503–527.

[Ya] Y. Yamada, *Stability of steady states for prey-predator diffusion equations with homogeneous Dirichlet conditions*, SIAM J. Math. Anal., 21 (1990), pp. 327–345.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 315–340

EXISTENCE OF SOLUTIONS AND A QUASI-STATIONARY LIMIT FOR A HYPERBOLIC SYSTEM DESCRIBING FERROMAGNETISM*

FRANK JOCHMANN†

Abstract. In this paper the transient Landau–Lifschitz equations coupled with Maxwell's equations describing ferromagnetic media without exchange interaction are studied. The main goals are existence of solutions and a quasi-stationary limit.

Key words. ferromagnetism, Maxwell's equations, existence of solutions, quasi-stationary limit

AMS subject classifications. 35Q60, 35L25, 78A40

PII. S0036141001392293

1. Introduction.

The paper is concerned with a nonlinear system consisting of Maxwell's equations

$$\varepsilon\partial_t\mathbf{E} = \operatorname{curl}\mathbf{H} - \sigma\mathbf{E} - \mathbf{J}, \qquad \mu\partial_t\mathbf{H} = -\operatorname{curl}\mathbf{E} - \mu\partial_t\tilde{\mathbf{M}}, \tag{1.1}$$

on $\mathbb{R}^+ \times \mathbb{R}^3$ coupled with the equation

$$\partial_t\mathbf{M} = F(x,\mathbf{M})\cdot\mathbf{H} + \mathbf{a}(x,\mathbf{M}) \tag{1.2}$$

on $\mathbb{R}^+ \times G$. Here $G \subset \mathbb{R}^3$ is an open set. In (1.1) the function $\tilde{\mathbf{M}}$ is the extension of $\mathbf{M}$ on $\mathbb{R}^+ \times \mathbb{R}^3$ defined by zero on the set $\mathbb{R}^+ \times (\mathbb{R}^3 \setminus G)$. This system, which describes the propagation of electromagnetic waves in a ferromagnetic medium occupying the set G, is supplemented by the initial conditions

$$\mathbf{E}(0,x) = \mathbf{E}_0(x), \qquad \mathbf{H}(0,x) = \mathbf{H}_0(x) \tag{1.3}$$

and

$$\mathbf{M}(0,x) = \mathbf{M}_0(x) \text{ on } G. \tag{1.4}$$

The unknown functions are the electric and magnetic field $\mathbf{E}, \mathbf{H}$, which depend on the time $t \geq 0$ and the space-variable $x \in \mathbb{R}^3$, and the magnetization $\mathbf{M}$ defined on $\mathbb{R}^+ \times G$; see [2], [13], [20].

Furthermore, $\varepsilon, \mu \in L^\infty(\mathbb{R}^3)$ denote the dielectric and magnetic permittivities, respectively, which are assumed to be bounded from below by a strictly positive constant. The electrical conductivity is denoted by σ for which the set $\{\sigma > 0\}$ does not necessarily coincide with G. An external current $\mathbf{J}$ is also included. The assumptions on the data $\mathbf{E}_0, \mathbf{H}_0$, and $\mathbf{J}$ are

$$\mathbf{J} \in L^1((0,\infty), L^2(\mathbb{R}^3)), \quad \mathbf{E}_0, \mathbf{H}_0 \in L^2(\mathbb{R}^3), \quad \mathbf{M}_0 \in L^2(G) \cap L^\infty(G). \tag{1.5}$$

*Received by the editors July 16, 2001; accepted for publication (in revised form) April 25, 2002; published electronically October 8, 2002. This research was partially supported by the Graduiertenkolleg *Analysis, Geometrie und ihre Verbindung zu den Naturwissenschaften* at the Mathematical Institute of the University of Leipzig, Germany.
http://www.siam.org/journals/sima/34-2/39229.html
†TU-Berlin, Fakultät II, Institut für Mathematik, Strasse D. 17 Juni 136, 10623 Berlin, Germany (jochmann@math.tu-berlin.de).

Furthermore, the initial state for the magnetic induction $\mathbf{B}_0 \stackrel{\text{def}}{=} \mu[\mathbf{H}_0+\tilde{\mathbf{M}}_0]$ is assumed to be divergence-free, i.e.,

$$\text{div } (\mu[\mathbf{H}_0 + \tilde{\mathbf{M}}_0]) = 0 \text{ on } \mathbb{R}^3. \tag{1.6}$$

It is assumed that the nonlinear functions $\mathbf{a} : G \times \mathbb{R}^3 \to \mathbb{R}^3$ and $F : G \times \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$ satisfy

$$\mathbf{ma}(x, \mathbf{m}) \leq 0 \text{ and } \mathbf{m}(F(x, \mathbf{m})\mathbf{h}) = 0 \text{ for all } x \in G, \mathbf{m} \in \mathbb{R}^3, \text{ and } \mathbf{h} \in \mathbb{R}^3. \tag{1.7}$$

Furthermore, they are assumed to be locally Lipschitz-continuous with respect to $\mathbf{M}$; i.e., for $A \in (0, \infty)$ there exists $L_A \in (0, \infty)$ such that

$$|\mathbf{a}(x, y) - \mathbf{a}(x, z)| + |F(x, y) - F(x, z)| \leq L_A|y - z| \tag{1.8}$$

for all $x \in G, y \in \mathbb{R}^3, \text{ and } z \in \mathbb{R}^3$ with $|y| + |z| \leq A$. Finally,

$$F(\cdot, 0) \in L^\infty(G) \text{ and } \mathbf{a}(\cdot, 0) \in L^2(G) \cap L^\infty(G). \tag{1.9}$$

Here $\mathbf{a}(x, \mathbf{M})$ takes into account a possible anisotropy of the medium, whereas $F(x, \mathbf{M})\mathbf{H}$ describes the interaction between the magnetic dipoles and the magnetic field $\mathbf{H}$. The dominant term is in most cases the torque $\mathbf{M} \wedge \mathbf{H}$, which is perpendicular to $\mathbf{M}$. This motivates assumption (1.7). A physically relevant example for F is

$$F(x, \mathbf{m})\mathbf{h} = -\gamma \mathbf{m} \wedge \mathbf{h} - \alpha|\mathbf{m}|^{-1}\mathbf{m} \wedge (\mathbf{m} \wedge \mathbf{h}) \tag{1.10}$$

including a damping term $\alpha\mathbf{m} \wedge (\mathbf{m} \wedge \mathbf{h})$ with $\alpha \geq 0$.

The main topic of this paper is a proof of existence of solutions $(\mathbf{E}, \mathbf{H}, \mathbf{M})$ of problem (1.1)–(1.4) and a quasi-stationary limit.

In the case $\varepsilon = \mu = 1$ on $\mathbb{R}^3$, the existence of solutions has been proved in [11], where the analysis is based on a compactness property of the commutator between P_H and smooth functions and also on a compensated compactness argument applied to the divergence-free part of $P_H(\mathbf{H}_n - \mathbf{H})$ for a suitable sequence $\{(\mathbf{E}_n, \mathbf{H}_n, \mathbf{M}_n)\}_{n\in\mathbb{N}}$ of approximate solutions. Here P_H denotes the orthogonal projector on the space of all vector fields $\mathbf{h} \in L^2(\mathbb{R}^3)$ with div $(\mu\mathbf{h}) = 0$. In the case $\varepsilon = \mu = 1$ on $\mathbb{R}^3$, Maxwell's equations for the part $P_H\mathbf{H}(\cdot)$ can be reduced to an inhomogeneous scalar wave equation and microlocal defect measures, and compensated compactness techniques (see [6], [16]) can be applied. This reduction to the scalar wave equation is no longer possible in the general case considered here, in which the coefficients ε, μ are not smooth, which often occurs in real situations. Therefore, a new compactness result for Maxwell's equations (Lemma 3.4) concerning local compactness properties of certain time averages of $P_H\mathbf{H}$ is proved in section 3. For the proof of the existence of solutions, in section 4 a sequence of approximate solutions $\{(\mathbf{E}_n, \mathbf{H}_n, \mathbf{M}_n)\}_{n\in\mathbb{N}}$ to a regularized problem is constructed. The subtle part is, due to the nonlinearity (1.2), the limit $n \to \infty$ where the abovementioned compactness result for Maxwell's equations and a lemma concerning the commutator between the projector P_H and suitable weight functions are used. The main ingredient of the proof of the existence of solutions is the strong convergence of $\{\mathbf{M}_n\}_{n\in\mathbb{N}}$ with respect to a weighted norm introduced in [11] in order to deal with the difficulty that $\mathbf{H}$ is generally not bounded. By a similar argument a weak convergence principle is obtained in section 5 which

says that the weak limit of solutions to (1.1)–(1.4) is again a solution provided that the initial data for $\mathbf{M}$ converge strongly.

Section 6 is concerned with the quasi-stationary limit which is of interest if the size of the ferromagnetic medium is very small in comparison to the electromagnetic wave length. After a suitable rescaling this corresponds to a high wave propagation speed $c = (\varepsilon\mu)^{-1/2}$. In this case one expects that the magnetic field is governed by magnetostatics, which means that

$$\text{curl } \mathbf{H} = \mathbf{J} \text{ and div } (\mu[\mathbf{H} + \tilde{\mathbf{M}}]) = 0 \text{ on } \mathbb{R}^+ \times \mathbb{R}^3.$$

For this purpose the external current $\mathbf{J}$ is assumed to be divergence-free. The precise statement concerning the quasi-stationary limit given in this paper is the following theorem.

THEOREM 1.1. *Assume* (1.5)–(1.9). *Suppose that in addition*

$$\mathbf{J} = \mathit{curl}\, \mathbf{g}_0 \tag{1.11}$$

with some $\mathbf{g}_0 \in L^2_{loc}(\mathbb{R}, H_{curl}) \cap W^{1,2}_{loc}(\mathbb{R}, L^2(\mathbb{R}^3))$, *in particular* $\operatorname{div} \mathbf{J} = 0$.

Let α_n *and* β_n *be sequences of positive numbers such that* $\alpha_n \overset{n\to\infty}{\longrightarrow} 0$, $\beta_n \overset{n\to\infty}{\longrightarrow} 0$, *and* α_n/β_n *is bounded as* $n \to \infty$. *With*

$$\varepsilon_n(x) = \alpha_n \varepsilon(x) \text{ and } \mu_n(x) = \beta_n \mu(x)$$

let $(\mathbf{E}_n, \mathbf{H}_n, \mathbf{M}_n)$ *be weak solutions of Landau–Lipschitz–Maxwell equations* (1.1)–(1.4), *where* ε *and* μ *are replaced by* ε_n *and* μ_n, *respectively.*

Then there exist $\mathbf{M} \in W^{1,\infty}_{loc}((0,\infty), L^2(G)) \cap L^\infty((0,\infty), L^\infty(G))$ *and a subsequence,* $(\mathbf{E}_{(n_m)}, \mathbf{H}_{(n_m)}, \mathbf{M}_{(n_m)})$, *such that for all* $T > 0$ *and* $p \in (2, \infty)$ *one has*

$$\mathbf{M}_{(n_m)} \overset{m\to\infty}{\longrightarrow} \mathbf{M} \text{ in } L^\infty((0,T), L^p(G)) \text{ strongly} \tag{1.12}$$

and

$$\mathbf{H}_{(n_m)} \overset{m\to\infty}{\longrightarrow} \mathbf{H} \text{ in } L^\infty((0,T), L^2(\mathbb{R}^3)) \text{ weak} * \tag{1.13}$$

with

$$\mathit{curl}\, \mathbf{H} = \mathbf{J} \text{ and } \operatorname{div}\left(\mu\left[\mathbf{H} + \tilde{\mathbf{M}}\right]\right) = 0. \tag{1.14}$$

Furthermore, $\mathbf{M}$ *and* $\mathbf{H}$ *solve*

$$\partial_t \mathbf{M} = F(x, \mathbf{M}) \cdot \mathbf{H} + \mathbf{a}(x, \mathbf{M}) \text{ on } \mathbb{R}^+ \times G \tag{1.15}$$

and

$$\mathbf{M}(0) = \mathbf{M}_0. \tag{1.16}$$

Here Lemma 3.4 is also applied. In the case where F is given by (1.10) with $\alpha = 0$, i.e., $F(x, \mathbf{m})\mathbf{h} = -\gamma \mathbf{m} \wedge \mathbf{h}$, the solution to problem (1.14)–(1.16) is unique and (1.12), (1.13) hold for the whole sequence.

Existence and the quasi-stationary limit have been carried out in [3] for the Landau–Lipschitz equation for the magnetic moment coupled with Maxwell's equations including the exchange interaction [3], [17], which reads as

$$\partial_t \mathbf{M} + \mathbf{M} \wedge \partial_t \mathbf{M} = 2\mathbf{M} \wedge (\Delta \mathbf{M} + \mathbf{H}) \text{ on } \mathbb{R}^+ \times G.$$

The parabolic structure of this equation simplifies the quasi-stationary limit considerably, since it provides an $H^1(G)$ estimate for the magnetic moment $\mathbf{M}$. This H^1 coercivity yields compactness properties of $\mathbf{M}$, which lets one pass to the limit in the nonlinear equations directly. Furthermore, it is shown in [3] and [4] that all points of the weak ω-limit set are solutions of the corresponding stationary equations. This is also a consequence of the H^1 boundedness of $\mathbf{M}$, which provides the compactness of the orbit with respect to the strong topology. The spatially one-dimensional case is studied in [12].

Other nonlinear models in electromagnetism have been studied, which involve the dielectric polarization $\mathbf{P}$ instead of the magnetic moment $\mathbf{M}$. In [8] and [10] the anharmonic oscillator model is studied, whereas in [5] and [9] the Maxwell–Bloch equations are considered.

2. Basic definitions and notation. Let $G \subset \mathbb{R}^3$ be a nonempty open set.

The dielectric and magnetic susceptibilities $\varepsilon, \mu \in L^\infty(\mathbb{R}^3)$ are assumed to be uniformly positive functions, which means that

$$\varepsilon(x), \mu(x) \geq a_0 \text{ on } \mathbb{R}^3 \text{ with some } a_0 > 0.$$

Furthermore, let $\sigma \in L^\infty(\mathbb{R}^3)$ be a nonnegative function.

Next, let H_{curl} be the space of all $\mathbf{E} \in L^2(\mathbb{R}^3, \mathbb{R}^3)$ with curl $\mathbf{E} \in L^2(\mathbb{R}^3)$. Now, the following operators are defined. As in [7] B is the skew self-adjoint operator

$$B(\mathbf{E}, \mathbf{H}) \stackrel{\text{def}}{=} (\varepsilon^{-1}\text{curl } \mathbf{H}, -\mu^{-1}\text{curl } \mathbf{E}) \text{ for } (\mathbf{E}, \mathbf{H}) \in D(B) \stackrel{\text{def}}{=} H_{curl} \times H_{curl}$$

in the Hilbert space $X \stackrel{\text{def}}{=} L^2(\mathbb{R}^3, \mathbb{C}^6)$ endowed with the scalar product

$$\langle (\mathbf{E}, \mathbf{H}), (\mathbf{F}, \mathbf{G}) \rangle_X \stackrel{\text{def}}{=} \int_{\mathbb{R}^3} \left(\varepsilon \mathbf{E} \cdot \overline{\mathbf{F}} + \mu \mathbf{H} \cdot \overline{\mathbf{G}} \right) dx.$$

The orthogonal decomposition (with respect to the L^2 scalar product without weight)

$$L^2(\mathbb{R}^3) = H_{curl,0} + H_{\text{div},0} \tag{2.1}$$

is well known, where $H_{curl,0}$ and $H_{\text{div},0}$ denote the spaces of all $\mathbf{E} \in L^2(\mathbb{R}^3, \mathbb{R}^3)$ with curl $\mathbf{E} = 0$ and div $\mathbf{E} = 0$, respectively. Let P with $P(\mathbf{e}, \mathbf{h}) = (P_E \mathbf{e}, P_H \mathbf{h})$ denote the orthogonal projector on $(\ker B)^\perp = \overline{\text{ran } B}$ (with respect to the weighted scalar product in X), which means that $(1 - P)$ is the orthogonal projector on $\ker B$. In particular,

$$\text{ran } (1 - P_E) = \text{ ran } (1 - P_H) = H_{curl,0}. \tag{2.2}$$

Since $(\mathbf{f}, \mathbf{g}) \in \ \ker B$ for all $\mathbf{f}, \mathbf{g} \in H_{curl,0}$, a pair $(\mathbf{e}, \mathbf{h}) \in L^2(\mathbb{R}^3)$ belongs to $(\ker B)^\perp = \text{ran } P$ if and only if

$$\int_{\mathbb{R}^3} (\varepsilon \mathbf{ef} + \mu \mathbf{hg}) dx = \langle (\mathbf{e}, \mathbf{h}), (\mathbf{f}, \mathbf{g}) \rangle_X = 0$$

for all $\mathbf{f}, \mathbf{g} \in H_{curl,0}$. By (2.1) this means that $\varepsilon \mathbf{e} \in H_{\text{div},0}$ and also $\mu \mathbf{h} \in H_{\text{div},0}$. Hence

$$\text{ran } P = (\ker B)^\perp = \{(\mathbf{e}, \mathbf{h}) \in L^2(\mathbb{R}^3) : \text{div } (\varepsilon \mathbf{e}) = \text{div } (\mu \mathbf{h}) = 0\} \tag{2.3}$$

in the sense of distributions. In particular,

$$\text{ran } P_H = \{\mathbf{h} \in L^2(\mathbb{R}^3) : \text{div } (\mu\mathbf{h}) = 0\}. \tag{2.4}$$

Next, let $F_\sigma : X \to X$ and $\mathcal{R} : L^2(G) \to X$ be defined by

$$(\mathcal{R}\mathbf{M})(x) \stackrel{\text{def}}{=} (0, \mathbf{M}(x)) \text{ if } x \in G \text{ and } (\mathcal{R}\mathbf{M})(x) \stackrel{\text{def}}{=} 0 \text{ if } x \in \mathbb{R}^3 \setminus G$$

and

$$F_\sigma(\mathbf{e}, \mathbf{h}) \stackrel{\text{def}}{=} \left(\varepsilon^{-1}\sigma\mathbf{e}, 0\right).$$

The aim of section 4 is to prove the existence of weak solutions with the properties

$$(\mathbf{E}, \mathbf{H}) \in C([0,\infty), X) \text{ and } \mathbf{M} \in W^{1,\infty}_{loc}([0,\infty), L^2(G)) \cap L^\infty_{loc}([0,\infty), L^\infty(G)). \tag{2.5}$$

This means that (1.1) is fulfilled in the sense that

$$\frac{d}{dt}\langle(\mathbf{E}(t), \mathbf{H}(t)), \mathbf{u}\rangle_X = -\langle(\mathbf{E}(t), \mathbf{H}(t)), B\mathbf{u}\rangle_X \tag{2.6}$$

$$- \langle \mathcal{R}\partial_t\mathbf{M}(t) + (\varepsilon^{-1}\mathbf{J}(t), 0) + F_\sigma(\mathbf{E}(t), \mathbf{H}(t)), \mathbf{u}\rangle_X \text{ for all } \mathbf{u} \in D(B).$$

This is equivalent to the variation of constant formula

$$(\mathbf{E}(t), \mathbf{H}(t)) = \exp(tB)\mathbf{w}_0 \tag{2.7}$$

$$- \int_0^t \exp((t-s)B)\left[\mathcal{R}\partial_t\mathbf{M}(s) + (\varepsilon^{-1}\mathbf{J}(s), 0) + F_\sigma(\mathbf{E}(s), \mathbf{H}(s))\right] ds,$$

where $\mathbf{w}_0 \stackrel{\text{def}}{=} (\mathbf{E}_0, \mathbf{H}_0) \in X$ and $\mathbf{J}$ are as in (1.5) and $(\exp(tB))_{t\in\mathbb{R}}$ is the unitary group generated by B; see [1], [7], and [14].

3. Some compactness results. One of the main goals of this section is to prove a generalization of the compensated compactness argument in [11] for Maxwell's equations with nonsmooth coefficients (Lemma 3.4). First it is shown that the space of all vector fields with divergence and curl in $L^q(\mathbb{R}^3)$, $q \in (6/5, 2]$, is compactly embedded in $L^2(B_r)$ for all $r > 0$, where $B_r \stackrel{\text{def}}{=} \{|x| < r\}$. Such a compactness result is well known for $q = 2$; see [15], [18], and [19]. This is generalized to the case where $\text{curl}\,\mathbf{h} \in L^q(\mathbb{R}^3)$ and $\text{div}\,(\mu\mathbf{h}) \in L^q(\mathbb{R}^3)$ for some $q \in (6/5, 2]$. Due to the fact that μ may be nonsmooth this does not follow directly from the Sobolev's compactness theorem $W^{1,q}(\mathbb{R}^3) \hookrightarrow L^2(B_r)$ for $q \in (6/5, 2]$.

Next, $H^{(q)}_{curl}$ denotes for $q \in [1, 2]$ the space of all $\mathbf{h} \in L^2(\mathbb{R}^3)$ with $\text{curl}\,\mathbf{h} \in L^q(\mathbb{R}^3) + L^2(\mathbb{R}^3)$; i.e., curl $\mathbf{h}$ admits a decomposition $\text{curl}\,\mathbf{h} = \mathbf{g}_1 + \mathbf{g}_2$ with

$$\mathbf{g}_1 \in L^q(\mathbb{R}^3) \text{ and } \mathbf{g}_2 \in L^2(\mathbb{R}^3).$$

The space $H^{(q)}_{\text{div}}$ is defined analogously. Now, a basic compactness lemma can be proved.

LEMMA 3.1. *Let $q \in (6/5, 2]$ and $r > 0$. Then the space of all $\mathbf{h} \in H^{(q)}_{curl}$ with $\mu\mathbf{h} \in H^{(q)}_{\text{div}}$ is compactly embedded in $L^2(B_r)$.*

Proof. Suppose that $\{\mathbf{h}_n\}_{n\in\mathbb{N}}$ is a bounded sequence in $H_{curl}(q)\cap\mu^{-1}H^{(q)}_{\mathrm{div}}$ which means that $\{\mathbf{h}_n\}_{n\in\mathbb{N}}$ is bounded in $L^2(\mathbb{R}^3)$, whereas $\{\mathrm{curl}\ \mathbf{h}_n\}_{n\in\mathbb{N}}$ and $\{\mathrm{div}\ [\mu\mathbf{h}_n]\}_{n\in\mathbb{N}}$ are bounded in $L^q(\mathbb{R}^3)+L^2(\mathbb{R}^3)$. Let

$$\chi\in C_0^\infty(B_{(2r)}) \text{ with } \chi=1 \text{ on } B_r \text{ and } \mathbf{U}_n \stackrel{\mathrm{def}}{=} \chi\mathbf{h}_n.$$

Then, since $q\le 2$ and supp χ is bounded, $\{\mathrm{curl}\ \mathbf{U}_n\}_{n\in\mathbb{N}}$ and $\{\mathrm{div}\ [\mu\mathbf{U}_n]\}_{n\in\mathbb{N}}$ are bounded in $L^q(\mathbb{R}^3)$ and supp $\mathbf{U}_n\subset B_{(2r)}$. Due to Poincaré's inequality

$$\|\psi\|^2_{H^1(B_{(2r)})}\le K\int_{B_{(2r)}}|\nabla\psi|^2dx$$

on the space

$$Y\stackrel{\mathrm{def}}{=}\left\{\psi\in H^1(B_{(2r)}) \text{ with } \int_{B_{(2r)}}\psi dx=0\right\}$$

there exists some $\psi_n\in Y$ with

$$\int_{B_{(2r)}}\mu\nabla\psi_n\nabla\psi dx=\int_{B_{(2r)}}\mu\mathbf{U}_n\nabla\psi dx \text{ for all } \psi\in Y \tag{3.1}$$

and, thus, for all $\psi\in H^1(B_{(2r)})$ (by adding a suitable constant). By taking $\psi=\psi_n$ in (3.1) it follows from the boundedness of $\mathbf{U}_n$ in $L^2(\mathbb{R}^3)$ that

$$\{\psi_n\}_{n\in\mathbb{N}} \text{ is bounded in } Y\subset H^1(B_{(2r)}). \tag{3.2}$$

Next, define $\mathbf{B}_n\in L^2(\mathbb{R}^3)$ by

$$\mathbf{B}_n(x)\stackrel{\mathrm{def}}{=}\mu(x)[\mathbf{U}_n(x)-\nabla\psi_n(x)] \text{ if } x\in B_{(2r)} \text{ and } \mathbf{B}_n(x)\stackrel{\mathrm{def}}{=}0 \text{ if } \in\mathbb{R}^3\setminus B_{(2r)}. \tag{3.3}$$

Then $\mathrm{div}\ \mathbf{B}_n=0$ on $\mathbb{R}^3$, since

$$\int_{\mathbb{R}^3}\mathbf{B}_n\nabla\varphi dx=\int_{B_{(2r)}}\mu(x)[\mathbf{U}_n(x)-\nabla\psi_n(x)]\nabla\varphi dx=0 \text{ for all } \varphi\in C_0^\infty(\mathbb{R}^3)$$

by (3.1). Furthermore, $\mathbf{B}_n$ is bounded in $L^2(\mathbb{R}^3)\cap L^1(\mathbb{R}^3)$, and hence $\widehat{\mathbf{B}_n}\in L^2(\mathbb{R}^3)\cap L^\infty(\mathbb{R}^3)$, where $\hat{\cdot}$ denotes Fourier-transform. Therefore,

$$\mathbf{g}_n(k)\stackrel{\mathrm{def}}{=}i|k|^{-2}k\wedge\widehat{\mathbf{B}_n}(k)\in L^2(\mathbb{R}^3).$$

Define $\mathbf{A}_n\stackrel{\mathrm{def}}{=}\mathcal{F}^{-1}\mathbf{g}_n\in L^2(\mathbb{R}^3)$. Since $\mathrm{div}\ \mathbf{B}_n=0$ one has $k\cdot\widehat{\mathbf{B}}_n(k)=0$ and thus $k\wedge\mathbf{g}_n(k)=-i\widehat{\mathbf{B}}_n(k)$. Hence

$$\mathrm{curl}\ \mathbf{A}_n=i\mathcal{F}^{-1}(k\wedge\mathbf{g}_n(k))=\mathbf{B}_n \text{ for all } n\in\mathbb{N}. \tag{3.4}$$

Since

$$\|\mathbf{A}_n\|_{H^1(\mathbb{R}^3)}=\|(1+|k|)\mathbf{g}_n\|_{L^2(\mathbb{R}^3)}\le C_1\|(|k|^{-1}+1)\widehat{\mathbf{B}_n}\|_{L^2(\mathbb{R}^3)}$$

$$\le C_2(\|\widehat{\mathbf{B}_n}\|_{L^\infty(\mathbb{R}^3)}+\|\widehat{\mathbf{B}_n}\|_{L^2(\mathbb{R}^3)})\le C_3(\|\mathbf{B}_n\|_{L^1(\mathbb{R}^3)}+\|\mathbf{B}_n\|_{L^2(\mathbb{R}^3)}),$$

it follows that

$$\{\mathbf{A}_n\}_{n\in\mathbb{N}} \text{ is bounded in } H^1(\mathbb{R}^3). \tag{3.5}$$

With supp $\mathbf{U}_n \subset B_{(2r)}$ one obtains from (3.3), (3.4), and Hölder's inequality after partial integration

$$c_0\|\mathbf{U}_n - \mathbf{U}_m\|^2_{L^2(B_{(2r)})} \le \int_{B_{(2r)}} (\mathbf{U}_n - \mathbf{U}_m)\cdot(\mu\mathbf{U}_n - \mu\mathbf{U}_m)dx$$

$$= \int_{B_{(2r)}} (\mathbf{U}_n - \mathbf{U}_m)(\text{curl } \mathbf{A}_n + \mu\nabla\psi_n - \text{curl } \mathbf{A}_m - \mu\nabla\psi_m)dx$$

$$= \int_{B_{(2r)}} [(\text{curl } \mathbf{U}_n - \text{curl } \mathbf{U}_m)(\mathbf{A}_n - \mathbf{A}_m) - \text{div } (\mu[\mathbf{U}_n - \mathbf{U}_m])(\psi_n - \psi_m)]dx$$

$$\le \|\text{curl } (\mathbf{U}_n - \mathbf{U}_m)\|_{L^q(B_{(2r)})}\|\mathbf{A}_n - \mathbf{A}_m\|_{L^{(q^*)}(B_{(2r)})}$$

$$+ \|\text{div } (\mu[\mathbf{U}_n - \mathbf{U}_m])\|_{L^q(B_{(2r)})}\|\psi_n - \psi_m\|_{L^{(q^*)}(B_{(2r)})}.$$

Since $q^* < 6$, the previous estimate, (3.2), (3.5), and Sobolev's embedding theorem $H^1(B_{2r}) \hookrightarrow L^{(q^*)}(B_{2r})$ yield the precompactness of $(\mathbf{U}_n)_{n\in\mathbb{N}}$ in $L^2(B_{2r})$, which completes the proof. □

Now, one obtains from Lemma 3.1, (2.2), and (2.4) the following corollary.

COROLLARY 3.2. *Let $q \in (6/5, 2]$ and $r > 0$. Then $(ran\ P_H) \cap H^{(q)}_{curl}$ is compactly embedded in $L^2(B_r)$.*

COROLLARY 3.3. *Let $\chi \in W^{1,\infty}(\mathbb{R}^3)$. Then the commutators $[\chi, P_E]$ and $[\chi, P_H]$ are compact operators from $L^2(\mathbb{R}^3)$ to $L^2(B_R)$ for all $R > 0$.*

Proof. Suppose $\mathbf{h} \in L^2(\mathbb{R}^3)$. Then $\mathbf{f} \overset{\text{def}}{=} \chi \cdot P_H\mathbf{h} - P_H(\chi\mathbf{h})$ obeys, by (2.4),

$$\text{div } (\mu\mathbf{f}) = \text{div } (\chi\mu P_H\mathbf{h}) = \mu\,(P_H\mathbf{h})\,\nabla\chi \in L^2(\mathbb{R}^3). \tag{3.6}$$

Since $\mathbf{f} = \chi\cdot(P_H - 1)\mathbf{h} - (P_H - 1)(\chi\mathbf{h})$, one obtains from (2.2)

$$\text{curl } \mathbf{f} = ((1 - P_H)\mathbf{h}) \wedge \nabla\chi \in L^2(\mathbb{R}^3).$$

By (3.6) and Lemma 3.1 this completes the proof. □

As a substitute for the compensated compactness argument in [11] the following compactness result for Maxwell's equations is proved.

LEMMA 3.4. *Suppose that $\{\mathbf{G}_n\}_{n\in\mathbb{N}}$, and $\{\mathbf{H}_n\}_{n\in\mathbb{N}}$ are bounded sequences in $L^\infty((0,T), L^2(\mathbb{R}^3))$, and $\{\mathbf{D}_n\}_{n\in\mathbb{N}}$ is bounded in $L^\infty((0,T), L^q(\mathbb{R}^3) + L^2(\mathbb{R}^3))$ with some $q \in (6/5, 2]$ such that*

$$\mathbf{H}_n \overset{n\to\infty}{\longrightarrow} 0 \text{ in } L^\infty((0,T), L^2(\mathbb{R}^3)) \text{ weak} * \tag{3.7}$$

and curl $\mathbf{H}_n(t) = \partial_t\mathbf{D}_n(t)$ on $(0,T)\times\mathbb{R}^3$ in the sense that

$$\int_{\mathbb{R}^3} \mathbf{H}_n(t)\cdot \text{curl } \mathbf{g}dx = \frac{d}{dt}\int_{\mathbb{R}^3} \mathbf{D}_n(t)\cdot\mathbf{g}dx \text{ for all } \mathbf{g} \in C_0^\infty(\mathbb{R}^3). \tag{3.8}$$

In addition, assume that the sequence $\{\mathbf{G}_n\}_{n\in\mathbb{N}}$ *is equicontinuous (from* $[0,T]$ *to* $L^2(\mathbb{R}^3)$*), i.e., for all* $\theta>0$ *there exists some* $\delta>0$ *such that*

(3.9) $\|\mathbf{G}_n(t)-\mathbf{G}_n(s)\|_{L^2(\mathbb{R}^3)}\leq\theta$ *for all* $n\in\mathbb{N}$ *and* $s,t\in(0,T)$ *with* $|s-t|\leq\delta$.

Then

$$\sup_{m\in\mathbb{N}}\int_0^T\int_{B_r}\mathbf{G}_m(t)\cdot P_H\mathbf{H}_n(t)dxdt\overset{n\to\infty}{\longrightarrow}0 \text{ for all radii } r>0.$$

Proof. The main ingredient of the proof is a local compactness property of certain time averages of $P_H\mathbf{H}_n$. For all $\chi\in C_0^\infty(0,T)$ one obtains from the boundedness of $\{\mathbf{G}_n\}_{n\in\mathbb{N}}$ and $\{P_H\mathbf{H}_n(\cdot)\}_{n\in\mathbb{N}}$ in $L^\infty((0,T),L^2(\mathbb{R}^3))$ that

$$\left|\int_0^T\int_{B_r}\mathbf{G}_m(t)\cdot P_H\mathbf{H}_n(t)dxdt-\int_0^T\int_{B_r}\chi(t)\mathbf{G}_m(t)\cdot P_H\mathbf{H}_n(t)dxdt\right|\leq C_1\int_0^T|1-\chi|dt$$

with some constant C_1 independent of $m,n\in\mathbb{N}$ and χ. Thus, it suffices to show for all $\chi\in C_0^\infty(0,T)$ and $r>0$

$$\sup_{m\in\mathbb{N}}\int_0^T\int_{B_r}\chi(t)\mathbf{G}_m(t)\cdot P_H\mathbf{H}_n(t)dxdt\overset{n\to\infty}{\longrightarrow}0. \tag{3.10}$$

Suppose $\chi\in C_0^\infty(0,T)$ and let

$$\mathbf{F}_m^{(h)}(t,x)\overset{\text{def}}{=}h^{-1}\int_0^h\chi(t+s)\mathbf{G}_m(t+s,x)ds \text{ for } h>0 \text{ and } m\in\mathbb{N}. \tag{3.11}$$

Since $\{\chi\mathbf{G}_n\}_{n\in\mathbb{N}}$ is also equicontinuous, one obtains

$$\sup_{t\in(0,T),m\in\mathbb{N}}\|\chi(t)\mathbf{G}_m(t)-\mathbf{F}_m^{(h)}(t)\|_{L^2(\mathbb{R}^3)}\overset{h\to0}{\longrightarrow}0,$$

and hence

$$\sup_{m,n\in\mathbb{N}}\left|\int_0^T\int_{B_r}\chi(t)\mathbf{G}_m(t)\cdot P_H\mathbf{H}_n(t)dxdt\right. \tag{3.12}$$

$$\left.-\int_0^T\int_{B_r}\chi(t)\mathbf{G}_m(t)\cdot P_H\mathbf{H}_n^{(h)}(t)dxdt\right|$$

$$=\sup_{m,n\in\mathbb{N}}\left|\int_0^T\int_{B_r}\Big(\chi(t)\mathbf{G}_m(t)-\mathbf{F}_m^{(h)}(t)\Big)\cdot P_H\mathbf{H}_n(t)dxdt\right|\overset{h\to0}{\longrightarrow}0,$$

where

$$\mathbf{H}_n^{(h)}(t,x)\overset{\text{def}}{=}h^{-1}\int_0^h\mathbf{H}_n(t-s,x)ds \text{ for } h>0,\ t\in\ \text{supp}\ \chi,\ \text{and } n\in\mathbb{N}. \tag{3.13}$$

Hence, it suffices to show that

$$\sup_{m\in\mathbb{N}}\int_0^T\int_{B_r}\chi(t)\mathbf{G}_m(t)\cdot P_H\mathbf{H}_n^{(h)}(t)dxdt\overset{n\to\infty}{\longrightarrow}0 \text{ for all } r>0, h>0. \tag{3.14}$$

Let $\mathbf{g}\in C_0^\infty(\mathbb{R}^3)$. Then it follows from (2.2) and (3.8) that

$$\int_{\mathbb{R}^3}P_H\mathbf{H}_n^{(h)}(t)\text{curl }\mathbf{g}dx=\int_{\mathbb{R}^3}\mathbf{H}_n^{(h)}(t)\text{curl }\mathbf{g}dx$$

$$=h^{-1}\int_0^h\int_{\mathbb{R}^3}\mathbf{H}_n(t-s)\text{curl }\mathbf{g}dxds=h^{-1}\int_{\mathbb{R}^3}[\mathbf{D}_n(t-h)-\mathbf{D}_n(t)]\cdot\mathbf{g}dx.$$

Hence

$$\text{curl }P_H\mathbf{H}_n^{(h)}(t)=h^{-1}[\mathbf{D}_n(t-h)-\mathbf{D}_n(t)]\in L^q(\mathbb{R}^3)+L^2(\mathbb{R}^3),$$

which implies by Corollary 3.2 and the boundedness of $\{\mathbf{D}_n\}_{n\in\mathbb{N}}$ in $L^\infty((0,T),L^q(\mathbb{R}^3)+L^2(\mathbb{R}^3))$ that

$$\{P_H\mathbf{H}_n^{(h)}(t)\}_{n\in\mathbb{N}}\text{ is precompact in }L^2(B_r)\text{ for fixed }t\in(0,T). \tag{3.15}$$

Since $\partial_t\mathbf{H}_n^{(h)}(t)=h^{-1}[\mathbf{H}_n(t)-\mathbf{H}_n(t-h)]$, it follows from the boundedness of $\{\mathbf{H}_n\}_{n\in\mathbb{N}}$ in $L^\infty((0,T),L^2(\mathbb{R}^3))$ that $\{P_H\mathbf{H}_n^{(h)}\}_{n\in\mathbb{N}}$ is bounded in $W^{1,\infty}((0,T),L^2(B_r))$. Hence it follows from (3.15) and Arzela's theorem that this sequence is precompact in $C([0,T],L^2(B_r))$ for all $r>0$. Thus, (3.7) yields

$$\|P_H\mathbf{H}_n^{(h)}(t)\|_{L^2(B_r)}\overset{n\to\infty}{\longrightarrow}0\text{ uniformly on }(0,T),$$

which gives (3.14), since $\{\mathbf{G}_n\}_{n\in\mathbb{N}}$ is bounded in $L^\infty((0,T),L^2(B_r))$. □

In what follows the linear, symmetric regularization operator $R_n: L^2(\mathbb{R}^3)\to L^2(\mathbb{R}^3)$ is defined by

$$(R_nf)(x)\overset{\text{def}}{=}\int_{\mathbb{R}^3}f(y)\omega_n(x-y)dy,\qquad x\in\mathbb{R}^3, \tag{3.16}$$

where $\omega_n\in C_0^\infty(\mathbb{R}^3)$ is a mollifier with supp $\omega_n\subset B_{(1/n)}$ and $\int_{\mathbb{R}^3}\omega_n dx=1$. Then R_n has the following properties:

$$\|F-R_nF\|_{L^2(\mathbb{R}^3)}\overset{n\to\infty}{\longrightarrow}0,\qquad \|R_nF\|_{L^2(\mathbb{R}^3)}\le K\|F\|_{L^2(\mathbb{R}^3)}, \tag{3.17}$$

$$\|R_nF\|_{L^2(B_r)}\le K\|F\|_{L^2(B_{(r+1)})}\text{ and }\|R_nF\|_{L^2(\mathbb{R}^3\setminus B_r)}\le K\|F\|_{L^2(\mathbb{R}^3\setminus B_{(r-1)})} \tag{3.18}$$

for all $r>1$ and $F\in L^2(\mathbb{R}^3)$ with some constant K independent of n, r, and F. For all $T>0$ and $f_\delta\in C_0^\infty((0,T)\times\mathbb{R}^3)$ the commutator between R_n and f_δ obeys

$$\sup_{t\in(0,T)}\|[f_\delta(t),R_n]\|_{B(L^2(\mathbb{R}^3),L^2(\mathbb{R}^3))}\overset{n\to\infty}{\longrightarrow}0. \tag{3.19}$$

For the proof of existence of solutions to (1.1)–(1.4) it is important that

$$R_nF\in L^\infty(\mathbb{R}^3)\text{ and }\|R_nF\|_{L^\infty(\mathbb{R}^3)}\le K_n\|F\|_{L^2(\mathbb{R}^3)}\text{ for all }F\in L^2(\mathbb{R}^3) \tag{3.20}$$

with some constant K_n independent of F but which may depend on n. Of course other regularizations with the properties (3.17)–(3.20) are possible. The following lemma concerns the commutator between the projector P_H and suitable weight functions. Beyond Lemma 3.4 it is also a main ingredient of the proof of the existence of solutions to (1.1)–(1.4) as well as of the proof of Theorem 1.1.

LEMMA 3.5. *Suppose that* $\{\mathbf{F}_n\}_{n\in\mathbb{N}}$ *is a bounded sequence in* $W^{1,\infty}((0,T),L^2(\mathbb{R}^3)) \cap L^\infty((0,T),L^\infty(\mathbb{R}^3))$ *with*

$$\mathbf{F}_n \overset{n\to\infty}{\longrightarrow} 0 \text{ in } L^\infty((0,T),L^2(\mathbb{R}^3)) \text{ weak } *. \tag{3.21}$$

Furthermore, let $\rho \in L^2((0,T),L^2(\mathbb{R}^3)) \cap L^\infty((0,T),L^\infty(\mathbb{R}^3))$. *Then*

$$\int_0^T \left\|\rho(t)^2 R_n(1-P_H)\mathbf{F}_n(t) - \rho(t)R_n(1-P_H)\left\{\rho(t)\mathbf{F}_n(t)\right\}\right\|_{L^1(\mathbb{R}^3)+L^2(\mathbb{R}^3)} dt \overset{n\to\infty}{\longrightarrow} 0$$

and

$$\int_0^T \left\|\rho(t)^2 (1-P_H)\mathbf{F}_n(t) - \rho(t)(1-P_H)\left\{\rho(t)\mathbf{F}_n(t)\right\}\right\|_{L^1(\mathbb{R}^3)+L^2(\mathbb{R}^3)} dt \overset{n\to\infty}{\longrightarrow} 0.$$

Proof. The idea is to approximate ρ by some smooth function $f \in C_0^\infty((0,T)\times\mathbb{R}^3)$ in order to get suitable estimates for the commutator between f and P_H and also for the commutator between f and R_n, since ρ may be nonsmooth.

For all $f \in C_0^\infty((0,T)\times\mathbb{R}^3)$ one obtains from the estimate in (3.17) and the boundedness of the sequence $\{\mathbf{F}_n\}_{n\in\mathbb{N}}$ in $L^\infty((0,T),L^2(\mathbb{R}^3)) \cap L^\infty((0,T),L^\infty(\mathbb{R}^3))$ that

$$\begin{aligned}
&\int_0^T \|\rho(t)\left[\rho(t)-f(t)\right] R_n(1-P_H)\mathbf{F}_n(t)\|_{L^1(\mathbb{R}^3)+L^2(\mathbb{R}^3)}\, dt \qquad (3.22)\\
&\leq \int_0^T \|\rho(t)\left[\rho(t)-f(t)\right] R_n(1-P_H)\mathbf{F}_n(t)\|_{L^1(\mathbb{R}^3)}\, dt\\
&\leq \int_0^T \|\rho(t)\|_{L^\infty(\mathbb{R}^3)} \|f(t)-\rho(t)\|_{L^2(\mathbb{R}^3)} \|R_n(1-P_H)\mathbf{F}_n(t)\|_{L^2(\mathbb{R}^3)}\, dt\\
&\leq C_1 \int_0^T \|f(t)-\rho(t)\|_{L^2(\mathbb{R}^3)} \|\mathbf{F}_n(t)\|_{L^2(\mathbb{R}^3)}\, dt\\
&\leq C_2 \|f-\rho\|_{L^2((0,T),L^2(\mathbb{R}^3))}
\end{aligned}$$

and also

$$\begin{aligned}
&\int_0^T \|\rho(t)R_n(1-P_H)\left\{\left[\rho(t)-f(t)\right]\mathbf{F}_n(t)\right\}\|_{L^1(\mathbb{R}^3)+L^2(\mathbb{R}^3)}\, dt \qquad (3.23)\\
&\leq C_3 \int_0^T \|(f(t)-\rho(t))\mathbf{F}_n(t)\|_{L^2(\mathbb{R}^3)}\, dt
\end{aligned}$$

$$\leq C_3 \int_0^T \|f(t)-\rho(t)\|_{L^2(\mathbb{R}^3)} \|\mathbf{F}_n(t)\|_{L^\infty(\mathbb{R}^3)}\, dt$$

$$\leq C_4 \|f-\rho\|_{L^2((0,T),L^2(\mathbb{R}^3))}$$

with some constants C_1, C_3 independent of n.

Let $\delta > 0$.

By the previous estimates there is some $f_\delta \in C_0^\infty((0,T)\times\mathbb{R}^3)$ such that

$$\int_0^T \big\|\rho(t)^2 R_n(1-P_H)\mathbf{F}_n(t) \tag{3.24}$$

$$-\ \rho(t)R_n(1-P_H)\{\rho(t)\mathbf{F}_n(t)\}\|_{L^1(\mathbb{R}^3)+L^2(\mathbb{R}^3)}\, dt$$

$$\leq \delta + \int_0^T \|\rho(t)f_\delta(t)R_n(1-P_H)\mathbf{F}_n(t) - \rho(t)R_n(1-P_H)\left\{f_\delta(t)\mathbf{F}_n(t)\right\}\|_{L^1(\mathbb{R}^3)+L^2(\mathbb{R}^3)}\, dt.$$

Since $f_\delta \in C_0^\infty((0,T)\times\mathbb{R}^3)$, it follows from (3.19) and the boundedness of the sequence $\{\mathbf{F}_n\}_{n\in\mathbb{N}}$ in $L^\infty((0,T),L^2(\mathbb{R}^3))$ again that we can choose some $N_\delta \in \mathbb{N}$ such that for all $n > N_\delta$

$$\int_0^T \|\rho(t)f_\delta(t)R_n(1-P_H)\mathbf{F}_n(t) \tag{3.25}$$

$$-\ \rho(t)R_n\left\{f_\delta(t)(1-P_H)\mathbf{F}_n(t)\right\}\|_{L^1(\mathbb{R}^3)+L^2(\mathbb{R}^3)}\, dt$$

$$\leq \int_0^T \|\rho(t)[f_\delta(t),R_n](1-P_H)\mathbf{F}_n(t)\|_{L^2(\mathbb{R}^3)}\, dt$$

$$\leq C_4 \|[f_\delta(t),R_n]\|_{B(L^2(\mathbb{R}^3),L^2(\mathbb{R}^3))} \leq \delta.$$

And thus, by (3.18) and (3.24), for all radii $r > 0$

$$\int_0^T \big\|\rho(t)^2 R_n(1-P_H)\mathbf{F}_n(t) \tag{3.26}$$

$$-\ \rho(t)R_n(1-P_H)\{\rho(t)\mathbf{F}_n(t)\}\|_{L^1(\mathbb{R}^3)+L^2(\mathbb{R}^3)}\, dt$$

$$\leq 2\delta + \int_0^T \|\rho(t)R_n\left\{f_\delta(t)(1-P_H)\mathbf{F}_n(t)\right\} - \rho(t)R_n(1-P_H)\left\{f_\delta(t)\mathbf{F}_n(t)\right\}\|_{L^1(\mathbb{R}^3)+L^2(\mathbb{R}^3)}\, dt$$

$$\leq 2\delta + \int_0^T \big(\|\rho(t)R_n\left\{[P_H,f_\delta(t)]\mathbf{F}_n(t)\right\}\|_{L^2(B_r)}$$

$$+\ \|\rho(t)R_n\left\{[P_H,f_\delta(t)]\mathbf{F}_n(t)\right\}\|_{L^1(\mathbb{R}^3\setminus B_r)}\big)\, dt$$

$$\leq 2\delta + C_5 \int_0^T (\|[P_H, f_\delta(t)]\mathbf{F}_n(t)\|_{L^2(B_{(r+1)})}$$

$$+ \|\rho(t)\|_{L^2(\mathbb{R}^3\setminus B_r)}\|R_n\left\{[P_H, f_\delta(t)]\mathbf{F}_n(t)\right\}\|_{L^2(\mathbb{R}^3)}\big)\, dt$$

$$\leq 2\delta + C_6 \int_0^T (\|[P_H, f_\delta(t)]\mathbf{F}_n(t)\|_{L^2(B_{(r+1)})} + \|\rho(t)\|_{L^2(\mathbb{R}^3\setminus B_r)})dt$$

with some constant C_6 independent of $n > N_\delta$ and r. Now, it follows from the boundedness of the sequence $\{\mathbf{F}_n\}_{n\in\mathbb{N}}$ in $L^\infty((0,T), L^2(\mathbb{R}^3))$ and Corollary 3.3 that

(3.27) $\{[P_H, f_\delta(t)]\mathbf{F}_n(t)\}_{n\in\mathbb{N}}$ is precompact in $L^2(B_{(r+1)})$ for fixed $t \in (0,T)$.

It follows from the boundedness of $\{\mathbf{F}_n\}_{n\in\mathbb{N}}$ in $W^{1,2}((0,T), L^2(\mathbb{R}^3))$ that the sequence $\{[P_H, f_\delta(\cdot)]\mathbf{F}_n(\cdot)\}_{n\in\mathbb{N}}$ is bounded in $W^{1,2}((0,T), L^2(B_{(r+1)}))$. Hence it follows from (3.27) and Arzela's theorem that this sequence is precompact in $C([0,T], L^2(B_{(r+1)}))$ for all $r > 0$. Thus, (3.21) yields

(3.28) $$\|[P_H, f_\delta(t)]\mathbf{F}_n(t)\|_{L^2(B_{(r+1)})} \overset{n\to\infty}{\longrightarrow} 0 \text{ uniformly on } (0,T)$$

for all $r > 0$. Now, (3.26) and (3.28) give

$$\limsup_{n\to\infty} \int_0^T \left\|\rho(t)^2 R_n(1-P_H)\mathbf{F}_n(t) - \rho(t)R_n(1-P_H)\left\{\rho(t)\mathbf{F}_n(t)\right\}\right\|_{L^1(\mathbb{R}^3)+L^2(\mathbb{R}^3)} dt$$

$$\leq 2\delta + C_6 \int_0^T \|\rho(t)\|_{L^2(\mathbb{R}^3\setminus B_r)} dt \text{ for all } r > 0.$$

Since $\delta > 0$ has been chosen arbitrarily, this completes the proof of the first assertion by letting $r \to \infty$. The other one is proved analogously. □

4. Existence of solutions. The main result of this section is the following theorem.

THEOREM 4.1. *Assume* (1.5)–(1.9). *Then problem* (1.1)–(1.4) *admits a weak solution* $(\mathbf{E}, \mathbf{H}, \mathbf{M})$ *with the properties* (2.5).

First a regularized problem is considered. Let

(4.1) $$\varepsilon\partial_t\mathbf{E}_n = \operatorname{curl}\mathbf{H}_n - \sigma\mathbf{E}_n - \mathbf{J}, \qquad \mu\partial_t\mathbf{H}_n = -\operatorname{curl}\mathbf{E}_n - \mu\partial_t\tilde{\mathbf{M}}_n,$$

on $\mathbb{R}^+ \times \mathbb{R}^3$, coupled with the equation

(4.2) $$\partial_t\mathbf{M}_n = F(x, \mathbf{M}_n)\cdot R_n(\mathbf{H}_n(\cdot)) + \mathbf{a}(x, \mathbf{M}_n)$$

on $\mathbb{R}^+ \times G$, with initial conditions (1.3) and (1.4). Here R_n is the regularization operator in (3.16). Due to the fact that R_n maps $L^2(\mathbb{R}^3)$ to $L^\infty(\mathbb{R}^3)$, problem (4.1)–(4.2) can be solved using the contraction mapping principle. The difficult part is the limit $n \to \infty$ where Lemmas 3.4 and 3.5 are used.

Suppose that $\mathbf{f} \in C_0(\mathbb{R}\times\mathbb{R}^3, \mathbb{R}^3)$ and let $\mathbf{m}$ be the solution to the ordinary initial value problem

(4.3) $$\partial_t\mathbf{m} = F(x, \mathbf{m})\cdot\mathbf{f} + \mathbf{a}(x, \mathbf{m})$$

(pointwise with respect to x) on $\mathbb{R}^+ \times G$ with initial condition (1.4). By assumption (1.7) multiplication with $\mathbf{m}$ gives $\mathbf{m}\partial_t\mathbf{m} \le 0$, and hence

$$|\mathbf{m}(t,x)| \le |\mathbf{M}_0(x)| \le C_0. \tag{4.4}$$

In particular, the ordinary initial value problem (4.3) and (1.4) admit a global solution $\mathbf{m} \in W^{1,\infty}_{loc}([0,\infty), L^2(G)) \cap L^\infty_{loc}([0,\infty), L^\infty(G))$ defined on $(0,\infty)\times G$.

Let $T > 0$ be arbitrary large and $\mathcal{A}_n : C([0,T],X) \to C([0,T],X)$ be defined by

$$(\mathcal{A}_n(\mathbf{E},\mathbf{H}))(t) = \exp(tB)(\mathbf{E}_0,\mathbf{H}_0)$$

$$-\int_0^t \exp((t-s)B)\left[\mathcal{R}\partial_t\mathbf{M}(s) + F_\sigma(\mathbf{E}(s),\mathbf{H}(s)) + (\varepsilon^{-1}\mathbf{J}(s),0)\right] ds,$$

where $\mathbf{M}$ solves (4.3) with $\mathbf{f}(t) \stackrel{\text{def}}{=} R_n(\mathbf{H}(t))$; i.e.,

$$\partial_t\mathbf{M} = F(x,\mathbf{M})\cdot R_n(\mathbf{H}_n(\cdot)) + \mathbf{a}(x,\mathbf{M}) \tag{4.5}$$

with initial condition (1.4). First, suppose that $(\mathbf{E},\mathbf{H}) \in C([0,T],X)$ and let $\mathbf{M} \in W^{1,2}([0,T], L^2(G,\mathbb{R}^3))$ be the solution to (4.5) and (1.4). Then $(\mathbf{E},\mathbf{H},\mathbf{M})$ solves (4.1), (4.2) on the interval $[0,T]$ with the initial conditions (1.3)–(1.4) (in the sense of (2.7)) if and only if $(\mathbf{E},\mathbf{H}) \in C([0,T],X)$ solves the fixed-point problem

$$\mathcal{A}_n(\mathbf{E},\mathbf{H}) = (\mathbf{E},\mathbf{H}). \tag{4.6}$$

Now suppose $(\mathbf{E},\mathbf{H}) \in C([0,T],X)$ and $(\mathbf{E}^{(1)},\mathbf{H}^{(1)}) \in C([0,T],X)$ and let $\mathbf{M}$ and $\mathbf{M}^{(1)}$ be the corresponding solutions to (4.5) and (1.4). With $R_n(\mathbf{H}(\cdot)) \in C([0,\infty), L^2(\mathbb{R}^3)\cap L^\infty(\mathbb{R}^3))$ by (3.20) one obtains from assumption (1.8), (1.9), the estimate in (3.17), and (4.4) that

$$\|\partial_t\mathbf{M}(t)\|_{L^2(G)} \le C_{1,n}(1 + \|R_n(\mathbf{H}(t))\|_{L^2(G)}) \le C_{2,n}(1 + \|(\mathbf{E}(t),\mathbf{H}(t))\|_X) \tag{4.7}$$

and

$$\begin{aligned}\|\partial_t\mathbf{M}(t) - \partial_t\mathbf{M}^{(1)}(t)\|_{L^2(G)} &\le C_{3,n}\|R_n(\mathbf{H}(t)) - R_n(\mathbf{H}^{(1)}(t))\|_{L^2(\mathbb{R}^3)} \\ &+ C_{3,n}\|\mathbf{M}(t) - \mathbf{M}^{(1)}(t)\|_{L^2(G)}\|R_n(\mathbf{H}(t))\|_{L^\infty(G)} + C_{3,n}\|\mathbf{M}(t) - \mathbf{M}^{(1)}(t)\|_{L^2(G)} \\ &\le C_{4,n}\|(\mathbf{E}(t),\mathbf{H}(t)) - (\mathbf{E}^{(1)}(t),\mathbf{H}^{(1)}(t))\|_X \\ &+ C_{4,n}\|\mathbf{M}(t) - \mathbf{M}^{(1)}(t)\|_{L^2(G)}(1 + \|(\mathbf{E}(t),\mathbf{H}(t))\|_X).\end{aligned} \tag{4.8}$$

The constants $C_{1,n} - C_{4,n}$ are independent of $(\mathbf{E},\mathbf{H})$, $(\mathbf{E}^{(1)},\mathbf{H}^{(1)})$, and t, but they may depend on n at this stage. Note that such an estimate generally does not hold for the original problem (1.1), (1.2) unless $\mathbf{H} \notin L^\infty_{loc}([0,\infty), L^\infty(\mathbb{R}^3))$. By (4.7), (4.8), and the standard energy estimate for weak solutions to the linear inhomogeneous Maxwell equations given by (2.7) it is now routine to show, by using the contraction mapping principle, that the fixed-point problem (4.6) has a unique fixed point $(\mathbf{E},\mathbf{H}) \in C([0,T],X)$. Hence problem (4.1), (4.2) has a unique solution on each finite

time interval $(0,T)$ and, therefore, it has a unique global solution $(\mathbf{E}_n, \mathbf{H}_n, \mathbf{M}_n)$ on $(0,\infty)$ the properties (2.5). It follows from (2.7) that

$$(1-P)(\mathbf{E}_n(t),\mathbf{H}_n(t)) = (1-P)(\mathbf{E}_0,\mathbf{H}_0) \tag{4.9}$$

$$-\int_0^t (1-P)\left[\mathcal{R}\partial_t \mathbf{M}_n(s) + (\varepsilon^{-1}\mathbf{J}(s),0) + F_\sigma(\mathbf{E}_n(s),\mathbf{H}_n(s))\right] ds.$$

In particular, by assumption (1.6) and (2.4),

$$(1-P_H)\left(\mathbf{H}_n(t)+\tilde{\mathbf{M}}_n(t)\right) = (1-P_H)\left(\mathbf{H}_0+\tilde{\mathbf{M}}_0\right) = 0, \tag{4.10}$$

where $\tilde{\mathbf{M}}_n(t)$ denotes the extension of $\mathbf{M}_n(t)$ by zero outside G. This means that the divergence-free condition on $\mathbf{B} \stackrel{\text{def}}{=} (\mu[\mathbf{H}+\tilde{\mathbf{M}}])$ is invariant under the nonlinear flow governed by (1.1), (1.2).

From now on the constants C_j are independent of $(\mathbf{E},\mathbf{H})$, t, and $n \in \mathbb{N}$. By (4.4) we have

$$\mathbf{M}_n \partial_t \mathbf{M}_n \le 0 \text{ and } |\mathbf{M}_n(t,x)| \le |\mathbf{M}_0(x)| \le C_0. \tag{4.11}$$

Now it follows from (4.2), (4.11), and the assumptions on the nonlinear functions that

$$|\partial_t \mathbf{M}_n| \le C_1 |R_n(\mathbf{H}_n(\cdot))| + C_1|\mathbf{M}_0|, \tag{4.12}$$

in particular, by the estimate in (3.17),

$$\|\mathcal{R}\partial_t \mathbf{M}_n(t)\|_X = \|\mu^{1/2}\partial_t \mathbf{M}_n(t)\|_{L^2(G)} \tag{4.13}$$

$$\le C_1(1+\|R_n(\mathbf{H}_n(t))\|_{L^2(G)}) \le C_2(1+\|(\mathbf{E}_n(t),\mathbf{H}_n(t))\|_X).$$

On the other hand, one obtains from (2.7) the energy estimate

$$\frac{1}{2}\frac{d}{dt}\|(\mathbf{E}_n(t),\mathbf{H}_n(t))\|_X^2 \le -\langle \mathcal{R}\partial_t \mathbf{M}_n(t) + (\varepsilon^{-1}\mathbf{J}(t),0), (\mathbf{E}_n(t),\mathbf{H}_n(t))\rangle_X \tag{4.14}$$

$$= -\int_G \mu \mathbf{H}_n(t)\partial_t \mathbf{M}_n(t) dx - \int_{\mathbb{R}^3} \mathbf{E}_n(t)\mathbf{J}(t) dx$$

$$\le \|(\mathbf{E}_n(t),\mathbf{H}_n(t))\|_X^2 + \|(\varepsilon^{-1}\mathbf{J}(t),0)\|_X^2 + \|\mathcal{R}\partial_t \mathbf{M}_n(t)\|_X^2.$$

By (4.11), (4.13), and (4.14) and Gronwall's lemma one obtains

$$\|(\mathbf{E}_n,\mathbf{H}_n)\|_{L^\infty((0,T),X)} + \|\mathbf{M}_n\|_{L^\infty((0,T),L^\infty(G))} \tag{4.15}$$

$$+ \|\partial_t \mathbf{M}_n\|_{L^\infty((0,T),L^2(G))} \le C_3$$

with some constants C_3 independent of n.

Hence there exists a subsequence $\{(\mathbf{E}_{n_m},\mathbf{H}_{n_m},\mathbf{M}_{n_m})\}_{m\in\mathbb{N}}$ such that

$$(\mathbf{E}_{n_m},\mathbf{H}_{n_m}) \stackrel{m\to\infty}{\longrightarrow} (\mathbf{E},\mathbf{H}) \text{ in } L^\infty((0,T),X) \text{ weak } *, \tag{4.16}$$

$$\mathbf{M}_{n_m} \overset{m\to\infty}{\longrightarrow} \mathbf{M} \text{ in } W^{1,2}((0,T),L^2(G)) \text{ weakly,} \tag{4.17}$$

and in $L^\infty((0,T),L^\infty(G))$ weak $*$. Note that it follows from the boundedness of $\partial_t \mathbf{M}_n$ in $L^2((0,T),L^2(G))$ and initial condition (1.4) for $\mathbf{M}_n$ that $\mathbf{M} \in W^{1,2}((0,T),L^2(G)) \subset C([0,T],L^2(G))$ and $\mathbf{M}(0) = \mathbf{M}_0$.

From (2.7), (4.1), (4.16), and (4.17) we obtain

$$(\mathbf{E},\mathbf{H}) = \exp(tB)(\mathbf{E}_0,\mathbf{H}_0) \tag{4.18}$$

$$-\int_0^t \exp((t-s)B)\left[\mathcal{R}\partial_t \mathbf{M}(s) + F_\sigma(\mathbf{E}(s),\mathbf{H}(s)) + (\varepsilon^{-1}\mathbf{J}(s),0)\right] ds;$$

i.e., $(\mathbf{E},\mathbf{H}) \in C([0,T],X)$ is the solution of the Maxwell system (1.1).

The aim of the following considerations is to show strong convergence of $\{\mathbf{M}_{n_m}\}_{m\in\mathbb{N}}$. By assumption (1.8) and the uniform boundedness of $\{\mathbf{M}_{n_m}\}_{m\in\mathbb{N}}$ in (4.11), there is some $L > 0$ such that

$$|F_n(t,x) - F_m(t,x)| \le L|\mathbf{M}_n(t,x) - \mathbf{M}_m(t,x)| \tag{4.19}$$

$$\text{and } |\mathbf{a}(x,\mathbf{M}_n) - \mathbf{a}(x,\mathbf{M}_m)| \le L|\mathbf{M}_n(t,x) - \mathbf{M}_m(t,x)| \text{ for all } n,m \in \mathbb{N}$$

with the abbreviation $F_n(t,x) \overset{\text{def}}{=} F(x,\mathbf{M}_n(t,x))$.

The main difficulty is that $\{\mathbf{H}_{n_m}\}_{m\in\mathbb{N}}$ is not uniformly bounded in general. For this purpose the weighted norm as in [11] is introduced. Let

$$\rho(t,x) \overset{\text{def}}{=} \rho_0(x)\exp\left(-L\int_0^t |\mathbf{H}(s,x)|ds\right) \text{ for } t \in (0,T) \tag{4.20}$$

with some arbitrarily chosen positive function $\rho_0 \in L^2(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3)$ and $\mathbf{H}$ as in (4.16). Then (4.19) gives

$$\int_G \rho(t)^2(\mathbf{M}_{n_m}(t) - \mathbf{M}_{n_p}(t)) \cdot \left(F_{n_m}(t) - F_{n_p}(t)\right)\mathbf{H}(t)dx \tag{4.21}$$

$$\le L\int_G \rho(t)^2|\mathbf{H}(t)|(\mathbf{M}_{n_m}(t) - \mathbf{M}_{n_p}(t))^2 dx.$$

Now, it follows from assumption (1.8), (4.2), (4.11), (4.20), and (4.21) that

$$\frac{1}{2}\frac{d}{dt}\|\rho(t)(\mathbf{M}_{n_m}(t) - \mathbf{M}_{n_p}(t))\|^2_{L^2(G)} \tag{4.22}$$

$$= \int_G \rho(t)^2(\mathbf{M}_{n_m}(t) - \mathbf{M}_{n_p}(t))\partial_t(\mathbf{M}_{n_m}(t) - \mathbf{M}_{n_p}(t))dx$$

$$- L\int_G \rho(t)^2|\mathbf{H}(t)|(\mathbf{M}_{n_m}(t) - \mathbf{M}_{n_p}(t))^2 dx$$

$$= \int_G \rho(t)^2(\mathbf{M}_{n_m}(t) - \mathbf{M}_{n_p}(t)) \cdot \left(F_{n_m}(t)R_{n_m}(\mathbf{H}_{n_m}(t)) - F_{n_p}(t)R_{n_p}(\mathbf{H}_{n_p}(t)\right) dx$$

$$+\int_G \rho(t)^2(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))\cdot\big(\mathbf{a}(x,\mathbf{M}_{n_m})-\mathbf{a}(x,\mathbf{M}_{n_p})\big)\,dx$$

$$-\ L\int_G \rho(t)^2|\mathbf{H}(t)|(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))^2dx$$

$$\leq \int_G \rho(t)^2(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))\cdot F_{n_m}(t)\left[R_{n_m}(\mathbf{H}_{n_m}(t))-\mathbf{H}(t)\right]dx$$

$$-\int_G \rho(t)^2(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))\cdot F_{n_p}(t)\left[R_{n_p}(\mathbf{H}_{n_p}(t))-\mathbf{H}(t)\right]dx$$

$$+\ C_3\|\rho(t)(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))\|^2_{L^2(G)}$$

$$\leq C_3\|\rho(t)(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))\|^2_{L^2(G)}+\sum_{j=1}^{3}h_{j,m,p}(t)+\sum_{j=1}^{3}h_{j,p,m}(t).$$

Here

$$h_{1,m,p}(t)\stackrel{\text{def}}{=}\int_G \rho(t)^2(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))\cdot F_{n_m}(t)R_{n_m}P_H\left(\mathbf{H}_{n_m}(t)-\mathbf{H}(t)\right)dx, \tag{4.23}$$

$$h_{2,m,p}(t)\stackrel{\text{def}}{=}\int_G \rho(t)^2(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))\cdot F_{n_m}(t)R_{n_m}(1-P_H)\left(\mathbf{H}_{n_m}(t)-\mathbf{H}(t)\right)dx, \tag{4.24}$$

$$h_{3,m,p}(t)\stackrel{\text{def}}{=}\int_G \rho(t)^2(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))\cdot F_{n_m}(t)\left[R_{n_m}(\mathbf{H}(t))-\mathbf{H}(t)\right]dx. \tag{4.25}$$

With (4.15) and the strong convergence in (3.17) it follows easily that

$$\int_0^T |h_{3,m,p}(t)|dt\stackrel{m,p\to\infty}{\longrightarrow}0. \tag{4.26}$$

In analogy to (4.10) it follows from (4.18) that

$$(1-P_H)\left(\mathbf{H}(t)+\tilde{\mathbf{M}}(t)\right)=(1-P_H)\left(\mathbf{H}_0+\tilde{\mathbf{M}}_0\right)=0, \tag{4.27}$$

and hence

$$(1-P_H)\left((\mathbf{H}_{n_m}(t))-\mathbf{H}(t)\right)=-(1-P_H)\left(\tilde{\mathbf{M}}_{n_m}(t)-\tilde{\mathbf{M}}(t)\right).$$

This gives

$$h_{2,m,p}(t) = \int_G \rho(t)^2(\mathbf{M}_{n_m}(t) - \mathbf{M}_{n_p}(t)) \cdot F_{n_m}(t) R_{n_m}(1-P_H)\left(\tilde{\mathbf{M}} - \tilde{\mathbf{M}}_{n_m}(t)\right) dx$$

$$= -\int_{\mathbb{R}^3} \left[\tilde{\mathbf{M}}_{n_m}(t) - \tilde{\mathbf{M}}_{n_p}(t)\right] \cdot \tilde{F}_{n_m}(t)\rho(t)^2 R_{n_m}(1-P_H)\left(\tilde{\mathbf{M}}(t) - \tilde{\mathbf{M}}_{n_m}(t)\right) dx.$$

The idea is to replace $\rho(t)R_{n_m}(1-P_H)(\tilde{\mathbf{M}}(t)-\tilde{\mathbf{M}}_{n_m}(t))$ by $R_{n_m}(1-P_H)\{\rho(t)(\tilde{\mathbf{M}}(t)-\tilde{\mathbf{M}}_{n_m}(t))\}$ in order to obtain an estimate of $h_{2,m,p}(t)$ in terms of the L^2-norm of $\rho(t)(\tilde{\mathbf{M}}(t) - \tilde{\mathbf{M}}_{n_m}(t))$.

Let $\delta > 0$.

By (4.15), (4.17), and Lemma 3.5 with

$$\mathbf{F}_m(t) \overset{\text{def}}{=} \tilde{\mathbf{M}}(t) - \tilde{\mathbf{M}}_{n_m}(t)$$

we can choose some $N_\delta \in \mathbb{N}$ such that for all $m, p > N_\delta$

$$\int_0^T \Big| h_{2,m,p}(t) - \int_{\mathbb{R}^3} \rho(t)\left[\tilde{\mathbf{M}}_{n_m}(t) - \tilde{\mathbf{M}}_{n_p}(t)\right] \tag{4.28}$$

$$\cdot\, \tilde{F}_{n_m}(t) R_{n_m}(1-P_H)\left\{\rho(t)\left(\tilde{\mathbf{M}}(t) - \tilde{\mathbf{M}}_{n_m}(t)\right)\right\} dx \Big| dt$$

$$\le \int_0^T \Big| \int_{\mathbb{R}^3} \tilde{F}_{n_m}(t)^* \left[\tilde{\mathbf{M}}_{n_m}(t) - \tilde{\mathbf{M}}_{n_p}(t)\right] \cdot \Big(\rho(t)^2 R_{n_m}(1-P_H)\left(\tilde{\mathbf{M}}(t) - \tilde{\mathbf{M}}_{n_m}(t)\right)$$

$$-\, \rho(t) R_{n_m}(1-P_H)\left\{\rho(t)\left(\tilde{\mathbf{M}}(t) - \tilde{\mathbf{M}}_{n_m}(t)\right)\right\}\Big) dx \Big| dt$$

$$\le \int_0^T \left\| \tilde{F}_{n_m}(t)^* \left[\tilde{\mathbf{M}}_{n_m}(t) - \tilde{\mathbf{M}}_{n_p}(t)\right] \right\|_{L^\infty(\mathbb{R}^3)\cap L^2(\mathbb{R}^3)}$$

$$\Big\| \rho(t)^2 R_{n_m}(1-P_H)\left(\tilde{\mathbf{M}}(t) - \tilde{\mathbf{M}}_{n_m}(t)\right)$$

$$-\rho(t) R_{n_m}(1-P_H)\left\{\rho(t)\left(\tilde{\mathbf{M}}(t) - \tilde{\mathbf{M}}_{n_m}(t)\right)\right\}\Big\|_{L^1(\mathbb{R}^3)+L^2(\mathbb{R}^3)} dt \le \delta,$$

and hence

$$\int_0^t |h_{2,m,p}(s)| ds \tag{4.29}$$

$$\le \delta + \int_0^t \Big| \int_{\mathbb{R}^3} \rho(s)[\tilde{\mathbf{M}}_{n_m}(s) - \tilde{\mathbf{M}}_{n_p}(s)]$$

$$\cdot\, \tilde{F}_{n_m}(s) R_{n_m}(1-P_H)\left\{\rho(s)\left(\tilde{\mathbf{M}}(s)-\tilde{\mathbf{M}}_{n_m}(s)\right)\right\}dx\Big|ds$$

$$\leq \delta + C_7 \int_0^t \left\|\rho(s)\left[\tilde{\mathbf{M}}_{n_p}(s)-\tilde{\mathbf{M}}_{n_m}(s)\right]\right\|_{L^2(\mathbb{R}^3)}$$

$$\left\|R_{n_m}(1-P_H)\left(\rho(s)\tilde{\mathbf{M}}_{n_m}(s)-\rho(s)\tilde{\mathbf{M}}(s)\right)\right\|_{L^2(\mathbb{R}^3)} ds$$

$$\leq \delta + C_8 \int_0^t \|\rho(s)(\mathbf{M}_{n_m}(s)-\mathbf{M}_{n_p}(s))\|_{L^2(G)} \|\rho(s)(\mathbf{M}_{n_m}(s)-\mathbf{M}(s))\|_{L^2(G)} ds$$

with some constant C_8 independent of t, δ and $m, p > N_\delta$.

It remains to estimate $h_{1,m,p}(t)$. Let

$$\mathbf{G}_{m,p}(t) \stackrel{\text{def}}{=} R_{n_m}\left\{\rho(t)^2 \tilde{F}_{n_m}(t)^*\left[\tilde{\mathbf{M}}_{n_m}(t)-\tilde{\mathbf{M}}_{n_p}(t)\right]\right\}. \tag{4.30}$$

By (4.23) one has

$$h_{1,m,p}(t) = \int_{\mathbb{R}^3} \mathbf{G}_{m,p}(t)\cdot P_H\left(\mathbf{H}_{n_m}(t)-\mathbf{H}(t)\right)dx$$

from which one obtains by (3.18) and (4.15) for all radii $r > 1$,

$$\left|\int_0^t h_{1,m,p}(s)ds\right| \leq \left|\int_0^t \int_{B_r} \mathbf{G}_{m,p}(s)\cdot P_H\left(\mathbf{H}_{n_m}(s)-\mathbf{H}(s)\right)dxds\right| \tag{4.31}$$

$$+ \int_0^T \|\mathbf{G}_{m,p}(s)\|_{L^2(\mathbb{R}^3\setminus B_r)} \|P_H\left(\mathbf{H}_{n_m}(s)-\mathbf{H}(s)\right)\|_{L^2(\mathbb{R}^3\setminus B_r)} ds$$

$$\leq \left|\int_0^t \int_{B_r} \mathbf{G}_{m,p}(s)\cdot P_H\left(\mathbf{H}_{n_m}(s)-\mathbf{H}(s)\right)dxds\right|$$

$$+ C_{10}\int_0^T \left\|\rho(s)^2 \tilde{F}_{n_m}(s)^*\left[\tilde{\mathbf{M}}_{n_m}(s)-\tilde{\mathbf{M}}_{n_p}(s)\right]\right\|_{L^2(\mathbb{R}^3\setminus B_{r-1})} ds$$

$$\leq \left|\int_0^t \int_{B_r} \mathbf{G}_{m,p}(s)\cdot P_H\left(\mathbf{H}_{n_m}(s)-\mathbf{H}(s)\right)dxds\right| + C_{11}\int_0^T \|\rho(s)^2\|_{L^2(\mathbb{R}^3\setminus B_{(r-1)})} ds$$

with some constant C_{11} independent of $m, p > N_\delta$ and r. By (4.1), (2.6), and (4.18), i.e., (1.1), one has

$$\int_{\mathbb{R}^3} (\mathbf{H}_{n_m}(t)-\mathbf{H}(t))\cdot \operatorname{curl}\mathbf{g}\,dx = \frac{d}{dt}\int_{\mathbb{R}^3} \mathbf{D}_{n_m}(t)\cdot\mathbf{g}\,dx \tag{4.32}$$

for all $\mathbf{g} \in C_0^\infty(\mathbb{R}^3)$ with

$$\mathbf{D}_n(t) \stackrel{\text{def}}{=} \varepsilon\left(\mathbf{E}_n(t)-\mathbf{E}(t)\right) + \int_0^t \sigma\left(\mathbf{E}_n(s)-\mathbf{E}(s)\right)ds.$$

By assumption (1.8), (1.9), (4.15), and (4.30) the functions $\{\tilde{\mathbf{M}}_{n_m}\}_{m\in\mathbb{N}}$ and $\{\rho^2\tilde{F}^*_{n_m}\}_{m\in\mathbb{N}}$ are bounded in $L^\infty((0,T),L^\infty(\mathbb{R}^3))$, whereas their time derivatives are bounded in $L^\infty((0,T),L^2(\mathbb{R}^3))$. Hence, by (3.17), there exists some constant K such that

$$\|\mathbf{G}_{m,p}(t)-\mathbf{G}_{m,p}(s)\|_{L^2(\mathbb{R}^3)} \le \Big\|\rho(t)^2\tilde{F}_{n_m}(t)^*\left[\tilde{\mathbf{M}}_{n_m}(t)-\tilde{\mathbf{M}}_{n_p}(t)\right]$$

$$-\,\rho(s)^2\tilde{F}_{n_m}(s)^*\left[\tilde{\mathbf{M}}_{n_m}(s)-\tilde{\mathbf{M}}_{n_p}(s)\right]\Big\|_{L^2(\mathbb{R}^3)} \le K|s-t| \text{ for all } s,t\in(0,T),$$

which means that the sequence $\{\mathbf{G}_{m,p}\}_{m\in\mathbb{N}}$ is equicontinuous. Therefore, it follows from (4.16), (4.32), and Lemma 3.4 that

$$\int_0^t\int_{B_r}\mathbf{G}_{m,p}(s)\cdot P_H\left(\mathbf{H}_{n_m}(s)-\mathbf{H}(s)\right)dxds \overset{m,p\to\infty}{\longrightarrow} 0 \text{ for all } r>1. \tag{4.33}$$

Now, (4.31) and (4.33) give

$$\limsup_{m,p\to\infty}\left|\int_0^t h_{1,m,p}(s)ds\right| \le C_{11}\int_0^T\|\rho(s)^2\|_{L^2(\mathbb{R}^3\setminus B_{(r-1)})}ds \text{ for all } r>1,$$

which implies that

$$\int_0^t h_{1,m,p}(s)ds \overset{m,p\to\infty}{\longrightarrow} 0 \text{ for all } t\in[0,T]. \tag{4.34}$$

It follows from (3.17), (4.15), and (4.23) that the functions $h_{1,m,p}$ are uniformly bounded, and hence the functions $\tilde{h}_{1,m,p}(t)\overset{\text{def}}{=}\int_0^t h_{1,m,p}(s)ds$ are equicontinuous on $[0,T]$. Therefore, the convergence in (4.34) is uniform with respect to $t\in[0,T]$. By (4.22), (4.26), (4.29), and (4.34) there is some $m_\delta>N_\delta$ such that

$$\frac{1}{2}\|\rho(t)(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))\|^2_{L^2(G)} \tag{4.35}$$

$$\le\sum_{j=1}^3\int_0^t h_{j,m,p}(s)ds+\sum_{j=1}^3\int_0^t h_{j,p,m}(s)ds$$

$$+\,C_3\int_0^t\|\rho(s)(\mathbf{M}_{n_m}(s)-\mathbf{M}_{n_p}(s))\|^2_{L^2(G)}ds$$

$$\le 6\delta+C_3\int_0^t\|\rho(s)(\mathbf{M}_{n_m}(s)-\mathbf{M}_{n_p}(s))\|^2_{L^2(G)}ds$$

$$+\,C_8\int_0^t\|\rho(s)(\mathbf{M}_{n_m}(s)-\mathbf{M}_{n_p}(s))\|_{L^2(G)}$$

$$\left(\|\rho(s)(\mathbf{M}_{n_m}(s)-\mathbf{M}(s))\|_{L^2(G)}+\|\rho(s)(\mathbf{M}_{n_p}(s)-\mathbf{M}(s))\|_{L^2(G)}\right)ds$$

$$\leq 6\delta + \frac{C_9}{6}\int_0^t (\|\rho(s)(\mathbf{M}_{n_m}(s)-\mathbf{M}_{n_p}(s))\|^2_{L^2(G)} + \|\rho(s)(\mathbf{M}_{n_m}(s)-\mathbf{M}(s))\|^2_{L^2(G)})ds$$

for all $t \in (0,T)$ and $m,p > m_\delta$. Using the elementary inequality

$$\exp(-C_9 t)C_9\int_0^t f(s)ds \leq \sup_{s\in(0,t)}[\exp(-C_9 s)f(s)] \leq \sup_{s\in(0,T)}[\exp(-C_9 s)f(s)]$$

for all nonnegative functions $f \in C[0,T]$ this gives

$$\sup_{t\in(0,T)}\left[\exp(-C_9 t)\|\rho(t)(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))\|^2_{L^2(G)}\right] \tag{4.36}$$

$$\leq 18\delta + \frac{1}{2}\sup_{t\in(0,T)}\left[\exp(-C_9 t)\|\rho(t)(\mathbf{M}_{n_m}(t)-\mathbf{M}(t))\|^2_{L^2(G)}\right].$$

Letting $p\to\infty$ it follows from (4.17) and (4.36) that

$$\exp(-C_9 t)\|\rho(t)(\mathbf{M}_{n_m}(t)-\mathbf{M}(t))\|^2_{L^2(G)} \tag{4.37}$$

$$\leq \limsup_{p\to\infty}\exp(-C_9 t)\|\rho(t)(\mathbf{M}_{n_m}(t)-\mathbf{M}_{n_p}(t))\|^2_{L^2(G)}$$

$$\leq 18\delta + \frac{1}{2}\sup_{t\in(0,T)}\left[\exp(-C_9 t)\|\rho(t)(\mathbf{M}_{n_m}(t)-\mathbf{M}(t))\|^2_{L^2(G)}\right].$$

This estimate gives

$$\|\rho[\mathbf{M}_{n_m}-\mathbf{M}]\|_{L^\infty((0,T),L^2(\mathbb{R}^3))} \overset{m\to\infty}{\longrightarrow} 0. \tag{4.38}$$

Let $A_k \overset{\text{def}}{=} \{(t,x)\in(0,T)\times G : \rho(t,x) > 1/k\}$. By (4.11) one obtains

$$\|\mathbf{M}_{n_m}-\mathbf{M}\|_{L^2((0,T),L^2(G))} \leq \|\mathbf{M}_{n_m}-\mathbf{M}\|_{L^2(A_k)} + \|2\mathbf{M}_0\|_{L^2([(0,T)\times G]\setminus A_k)}$$

$$\leq k\|\rho[\mathbf{M}_{n_m}-\mathbf{M}]\|_{L^2((0,T),L^2(\mathbb{R}^3))} + \|2\mathbf{M}_0\|_{L^2([(0,T)\times G]\setminus A_k)}.$$

Now (4.38) yields

$$\limsup_{m\to\infty}\|\mathbf{M}_{n_m}-\mathbf{M}\|_{L^2((0,T),L^2(\mathbb{R}^3))} \leq \|2\mathbf{M}_0\|_{L^2([(0,T)\times G]\setminus A_k)} \text{ for all } k\in\mathbb{N}.$$

Since $(0,T)\times G = \bigcup_{k=1}^\infty A_k$, this gives

$$\|\mathbf{M}_{n_m}-\mathbf{M}\|_{L^2((0,T),L^2(\mathbb{R}^3))} \overset{m\to\infty}{\longrightarrow} 0. \tag{4.39}$$

By (3.17), (4.2), and (4.16) it follows from this strong convergence that $(\mathbf{E},\mathbf{H},\mathbf{M})$ also satisfies (1.2). Since T can be chosen arbitrarily large this completes the proof of the existence of solutions.

5. A weak convergence principle. The following theorem says that the weak limit of solutions to (1.1)–(1.4) is again a solution provided that the initial data for $\mathbf{M}$ converge strongly. It will also be used in section 6.

THEOREM 5.1. *Assume* (1.7)–(1.9). *Suppose that* $\{\mathbf{H}_n\}_{n\in\mathbb{N}}$ *is a bounded sequence in* $L^\infty((0,T),L^2(\mathbb{R}^3))$, *and let* $\{\mathbf{M}_n\}_{n\in\mathbb{N}}$ *be a bounded sequence in* $W^{1,\infty}((0,T),L^2(G)) \cap L^\infty((0,T),L^\infty(G))$, *such that*

$$\mathbf{H}_n \overset{n\to\infty}{\longrightarrow} \mathbf{H} \text{ in } L^\infty((0,T),L^2(\mathbb{R}^3)) \text{ weak } *, \tag{5.1}$$

$$\mathbf{M}_n \overset{n\to\infty}{\longrightarrow} \mathbf{M} \text{ in } L^\infty((0,T),L^\infty(G)) \text{ weak } *, \tag{5.2}$$

and

$$\mathbf{M}_n(0) \overset{n\to\infty}{\longrightarrow} \mathbf{M}_0 \text{ in } L^2(G) \text{ strongly }. \tag{5.3}$$

Furthermore, assume that

$$(1-P_H)\mathbf{H}_n(t) = (1-P_H)\tilde{\mathbf{M}}_n(t), \tag{5.4}$$

$$\partial_t \mathbf{M}_n = F(x,\mathbf{M}_n)\cdot\mathbf{H}_n + \mathbf{a}(x,\mathbf{M}_n) \text{ on } \mathbb{R}^+\times G, \tag{5.5}$$

and

$$\int_{\mathbb{R}^3} \mathbf{H}_n(t)\cdot \mathit{curl}\,\mathbf{g}dx = \frac{d}{dt}\int_{\mathbb{R}^3}\mathbf{D}_n(t)\cdot\mathbf{g}dx \text{ for all } \mathbf{g}\in C_0^\infty(\mathbb{R}^3), \tag{5.6}$$

where $\{\mathbf{D}_n\}_{n\in\mathbb{N}}$ *is a bounded sequence in* $L^\infty((0,T),L^q(\mathbb{R}^3)+L^2(\mathbb{R}^3))$ *for some* $q\in(6/5,2]$. *Then*

$$\|\mathbf{M}_n-\mathbf{M}\|_{L^\infty((0,T),L^p(G))} \overset{n\to\infty}{\longrightarrow} 0 \text{ for all } p\in[2\infty), \tag{5.7}$$

$$\partial_t\mathbf{M} = F(x,\mathbf{M})\cdot\mathbf{H} + \mathbf{a}(x,\mathbf{M}) \text{ on } \mathbb{R}^+\times G, \tag{5.8}$$

and

$$\mathbf{M}(0) = \mathbf{M}_0. \tag{5.9}$$

Proof. The basic idea is to show strong convergence of $\{\mathbf{M}_{n_m}\}_{m\in\mathbb{N}}$ by using similar arguments as in the proof of Theorem 4.1.

Let $F_n(t,x) \overset{\text{def}}{=} F(x,\mathbf{M}_n(t))$ and ρ be as in (4.20) with $\mathbf{H}$ as in (5.1). As in (4.22) one obtains

$$\begin{aligned}\frac{1}{2}\|\rho(t)(\mathbf{M}_n(t)-\mathbf{M}_m(t))\|^2_{L^2(G)} &\le \frac{1}{2}\|\rho(t)(\mathbf{M}_n(0)-\mathbf{M}_m(0))\|^2_{L^2(G)} \\ &+ C_1\int_0^t \|\rho(s)(\mathbf{M}_n(s)-\tilde{\mathbf{M}}_m(s))\|^2_{L^2(G)}ds \\ &+ \int_0^t (g_{1,n,m}(s)+g_{2,n,m}(s)+g_{1,m,n}(s)+g_{2,m,n}(s))\,ds.\end{aligned} \tag{5.10}$$

Here

$$g_{1,n,m}(t) \stackrel{\text{def}}{=} \int_G \rho(t)^2[\mathbf{M}_n(t)-\mathbf{M}_m(t)]\cdot F_n(t)P_H\left(\mathbf{H}_n(t)-\mathbf{H}(t)\right)dx, \tag{5.11}$$

$$\begin{aligned} g_{2,n,m}(t) &\stackrel{\text{def}}{=} \int_G \rho(t)^2[\mathbf{M}_n(t)-\mathbf{M}_m(t)]\cdot F_n(t)(1-P_H)\left(\mathbf{H}_n(t)-\mathbf{H}(t)\right)dx \\ &= -\int_{\mathbb{R}^3} \rho(t)\left[\tilde{\mathbf{M}}_n(t)-\tilde{\mathbf{M}}_m(t)\right]\cdot \tilde{F}_n(t)\rho(t)(1-P_H)\left(\tilde{\mathbf{M}}(t)-\tilde{\mathbf{M}}_n(t)\right)dx, \end{aligned} \tag{5.12}$$

by 5.4, where $\tilde{\mathbf{M}}(t)$ denotes the extension of $\mathbf{M}(t)$ by zero outside G. The estimate of $g_{2,n,m}(t)$ can be arranged as in the proof of Theorem 4.1 using (5.2) and Lemma 3.5.

In order to estimate $g_{1,n,m}(t)$ let

$$\mathbf{G}_{n,m}(t) \stackrel{\text{def}}{=} \rho(t)^2\tilde{F}_n(t)^*\left[\tilde{\mathbf{M}}_n(t)-\tilde{\mathbf{M}}_m(t)\right]. \tag{5.13}$$

By (5.11) one has

$$g_{1,n,m}(t) = \int_{\mathbb{R}^3} \mathbf{G}_{n,m}(t)\cdot P_H\left(\mathbf{H}_n(t)-\mathbf{H}(t)\right)dx.$$

In analogy to (4.31) and (4.33) it suffices to show that

$$\int_0^t \int_{B_r} \mathbf{G}_{n,m}(s)\cdot P_H\left(\mathbf{H}_n(s)-\mathbf{H}(s)\right)dxds \stackrel{m,n\to\infty}{\longrightarrow} 0 \tag{5.14}$$

for all $t\in[0,T]$ and $r>0$.

It follows from the boundedness of $\{\mathbf{D}_n\}_{n\in\mathbb{N}}$ there exists some $\mathbf{D}\in L^2((0,T),L^q(\mathbb{R}^3)+L^2(\mathbb{R}^3))$ and a subsequence $\{\mathbf{D}_{n_m}\}_{m\in\mathbb{N}}$ such that

$$\mathbf{D}_{n_m} \stackrel{m\to\infty}{\longrightarrow} \mathbf{D} \text{ in } L^\infty((0,T),L^q(\mathbb{R}^3)+L^2(\mathbb{R}^3)) \text{ weak } -*. \tag{5.15}$$

By (5.1), (5.6), and (5.15) one obtains

$$\int_{\mathbb{R}^3} \mathbf{H}(t)\cdot \operatorname{curl}\mathbf{g}dx = \frac{d}{dt}\int_{\mathbb{R}^3}\mathbf{D}(t)\cdot\mathbf{g}dx,$$

and hence, by (5.6),

$$\int_{\mathbb{R}^3} (\mathbf{H}_n(t)-\mathbf{H}(t))\cdot\operatorname{curl}\mathbf{g}dx = \frac{d}{dt}\int_{\mathbb{R}^3}(\mathbf{D}_n(t)-\mathbf{D}(t))\cdot\mathbf{g}dx \tag{5.16}$$

for all $\mathbf{g}\in C_0^\infty(\mathbb{R}^3)$. From (1.8), (1.9), and the boundedness of $\{\mathbf{M}_n\}_{n\in\mathbb{N}}$ in

$$W^{1,\infty}((0,T),L^2(G))\cap L^\infty((0,T),L^\infty(G))$$

one obtains the equicontinuity of the family $\{\mathbf{G}_{n,m}\}_{n,m\in\mathbb{N}}$ required for Lemma 3.4. Hence (5.14) follows from (5.1), (5.16), and Lemma 3.4. Proceeding as in the proof of Theorem 4.1 one gets (5.7) for $p=2$. By the uniform boundedness of $\{\mathbf{M}_n\}_{n\in\mathbb{N}}$ (5.7) also holds for all $p\in[2,\infty)$. By (5.1), (5.5) it follows from this strong convergence that $(\mathbf{E},\mathbf{H},\mathbf{M})$ satisfies (5.8). □

6. The quasi-stationary limit. In what follows let α_n and β_n be sequences of positive numbers with

$$\alpha_n \overset{n\to\infty}{\longrightarrow} 0,\ \beta_n \overset{n\to\infty}{\longrightarrow} 0,\ \text{and } \alpha_n/\beta_n \le K$$

with some constant K independent of n. Furthermore, let $(\mathbf{E}_n, \mathbf{H}_n, \mathbf{M}_n)$ be a weak solution to (1.1)–(1.4) where ε and μ are replaced by $\varepsilon_n = \alpha_n\varepsilon$ and $\mu_n = \beta_n\mu$, respectively. Setting $\mathbf{h}_n \overset{\text{def}}{=} \mathbf{H}_n - \mathbf{g}_0$ with $\mathbf{g}_0$ as in (1.11) these equations read as

$$\alpha_n\varepsilon\partial_t\mathbf{E}_n = \varepsilon_n\partial_t\mathbf{E}_n = \text{curl } \mathbf{H}_n - \sigma\mathbf{E}_n - \mathbf{J} = \text{curl } \mathbf{h}_n - \sigma\mathbf{E}_n, \tag{6.1}$$

$$\beta_n\mu\partial_t\mathbf{h}_n = \mu_n\partial_t\mathbf{H}_n - \beta_n\mu\partial_t\mathbf{g}_0 = -\text{curl } \mathbf{E}_n - \beta_n\mu\partial_t\tilde{\mathbf{M}}_n - \beta_n\mu\partial_t\mathbf{g}_0, \tag{6.2}$$

on $\mathbb{R}^+ \times \mathbb{R}^3$ coupled with the equation

$$\partial_t\mathbf{M}_n = F(x, \mathbf{M}_n)\cdot\mathbf{H}_n + \mathbf{a}(x, \mathbf{M}_n) \tag{6.3}$$

on $\mathbb{R}^+ \times G$, with initial conditions (1.3) and (1.4).

Proof of Theorem 1.1. Let $T > 0$ be arbitrary large. From (6.1) and (6.2) one obtains the energy balance

$$\frac{1}{2}\frac{d}{dt}\|(\alpha_n^{1/2}\mathbf{E}_n(t), \beta_n^{1/2}\mathbf{h}_n(t))\|_X^2 \tag{6.4}$$

$$= -\int_{\mathbb{R}^3}\mathbf{E}_n(t)\sigma\mathbf{E}_n(t)dx - \beta_n\int_G\mu\mathbf{h}_n(t)\partial_t\mathbf{M}(t)dx - \beta_n\int_{\mathbb{R}^3}\mu\mathbf{h}_n(t)\partial_t\mathbf{g}_0(t)dx.$$

This gives

$$\frac{1}{2}\|(\alpha_n^{1/2}\beta_n^{-1/2}\mathbf{E}_n(t), \mathbf{h}_n(t))\|_X^2 \tag{6.5}$$

$$\le \frac{1}{2}\|(\alpha_n^{1/2}\beta_n^{-1/2}\mathbf{E}_0, \mathbf{h}_0)\|_X^2 - \beta_n^{-1}\int_0^t\int_{\mathbb{R}^3}\mathbf{E}_n(s)\sigma\mathbf{E}_n(s)dxds$$

$$+\int_0^t\left(\|\mu^{1/2}\mathbf{h}_n(s)\|_{L^2(\mathbb{R}^3)}^2 + \|\mu^{1/2}\partial_t\mathbf{M}_n(s)\|_{L^2(G)}^2 + \|\mu^{1/2}\partial_t\mathbf{g}_0(s)\|_{L^2(\mathbb{R}^3)}^2\right)ds.$$

From (1.5), (4.4) one obtains the following bound on $\mathbf{M}_n$:

$$\|\mathbf{M}_n\|_{L^\infty((0,T),L^\infty(G))} + \|\mathbf{M}_n\|_{L^\infty((0,T),L^2(G))} \le C_1. \tag{6.6}$$

By assumption (1.8), (1.9), (6.3), and (6.6) one obtains

$$\int_0^t\|\mu^{1/2}\partial_t\mathbf{M}_n(s)\|_{L^2(G)}^2ds \le C_2\int_0^t(1 + \|F(x, \mathbf{M}_n(s))\|_{L^\infty(G)}\|\mu^{1/2}\mathbf{H}_n(s)\|_{L^2(\mathbb{R}^3)})^2ds$$

$$\le C_3\left(1 + \int_0^t\|\mu^{1/2}\mathbf{h}_n(s)\|_{L^2(\mathbb{R}^3)}^2ds\right).$$

Next, (6.5) and this estimate yield $\|(\alpha_n^{1/2}\beta_n^{-1/2}\mathbf{E}_n(t), \mathbf{h}_n(t))\|_X^2 \leq C_4$, and hence

$$\alpha_n \|\mathbf{E}_n\|_{L^\infty((0,T),L^2(\mathbb{R}^3))} \leq C_5 \alpha_n^{1/2}\beta_n^{1/2}\,, \qquad \|\mathbf{H}_n\|_{L^\infty((0,T),L^2(\mathbb{R}^3))} \leq C_5 \tag{6.7}$$

and

$$\|\sigma \mathbf{E}_n\|_{L^2((0,T),L^2(\mathbb{R}^3))} \leq C_5 \beta_n^{1/2}.$$

Furthermore, assumption (1.8), (1.9), (6.3), (6.6), and (6.7) also give

$$\|\partial_t \mathbf{M}_n\|_{L^\infty((0,T),L^2(G))} \leq C_5 \tag{6.8}$$

with some constant C_5 independent of n.

By (6.6) and (6.7) there exists a subsequence $\{(\mathbf{E}_{n_m}, \mathbf{H}_{n_m}, \mathbf{M}_{n_m})\}_{m\in\mathbb{N}}$ such that

$$\mathbf{H}_{n_m} \overset{m\to\infty}{\longrightarrow} \mathbf{H} \text{ in } L^\infty((0,T),L^2(\mathbb{R}^3)) \text{ weak } *, \tag{6.9}$$

$$\mathbf{M}_{n_m} \overset{m\to\infty}{\longrightarrow} \mathbf{M} \text{ in } L^\infty((0,T),L^\infty(G)) \text{ weak } *, \tag{6.10}$$

and in $W^{1,2}((0,T),L^2(G))$ weakly.

By (2.4), assumption (1.6), (2.6), and (6.2) one has

$$(1-P_H)\left(\mathbf{H}_{n_m}(t) + \tilde{\mathbf{M}}_{n_m}(t)\right) = (1-P_H)\left(\mathbf{H}_0 + \tilde{\mathbf{M}}_0\right) = 0. \tag{6.11}$$

Next, (6.1) and (6.7) imply that

$$\operatorname{curl} \mathbf{H}_n - \mathbf{J} = \alpha_n \varepsilon \partial_t \mathbf{E}_n + \sigma \mathbf{E}_n \overset{n\to\infty}{\longrightarrow} 0 \text{ in } \mathcal{D}'\left((0,\infty)\times\mathbb{R}^3\right).$$

Hence, one obtains from (6.9)–(6.11) that

$$\operatorname{curl} \mathbf{H} = \mathbf{J} \text{ and } \operatorname{div}\left(\mu\left[\mathbf{H} + \tilde{\mathbf{M}}\right]\right) = 0 \text{ on } (0,\infty)\times\mathbb{R}^3. \tag{6.12}$$

Now, Theorem 5.1 will be applied. By (6.9), (6.10), and (6.11) the conditions (5.1)–(5.4) are fulfilled, and by (6.1) and (2.6) one has

$$\int_{\mathbb{R}^3} \mathbf{H}_{n_m}(t)\cdot \operatorname{curl} \mathbf{g}\, dx = \frac{d}{dt}\int_{\mathbb{R}^3} \mathbf{D}_{n_m}(t)\cdot \mathbf{g}\, dx \text{ for all } \mathbf{g}\in C_0^\infty(\mathbb{R}^3),$$

with

$$\mathbf{D}_n(t) \overset{\text{def}}{=} \alpha_n \varepsilon \mathbf{E}_n(t) + \int_0^t (\sigma \mathbf{E}_n(s) + \mathbf{J}(s))\, ds,$$

which is bounded in $L^\infty((0,T),L^2(\mathbb{R}^3))$ by (6.7). Thus, assumption (5.6) is also satisfied. Finally, the assertion follows from (6.12) and Theorem 5.1. □

Remark 1. By (2.2), (2.4), and (1.11) it follows that (1.14) is fulfilled if and only if

$$\mathbf{H}(t) \overset{\text{def}}{=} P_H \mathbf{g}_0(t) - (1-P_H)\tilde{\mathbf{M}}(t) \text{ for all } t\in(0,\infty). \tag{6.13}$$

LEMMA 6.1. *Suppose that in addition F is given by*

$$F(x,\mathbf{m})\mathbf{h} = -\gamma(x)\mathbf{m}\wedge\mathbf{h} \text{ for all } x\in G, \mathbf{m}\in\mathbb{R}^3 \text{ and } \mathbf{h}\in\mathbb{R}^3 \tag{6.14}$$

with some function $\gamma\in L^\infty(G)$. The solution to problem (1.14)–(1.16) *is unique and* (1.12), (1.13) *hold for the whole sequence.*

Proof. Suppose that $(\mathbf{H}^{(1)},\mathbf{M}^{(1)})$ and $(\mathbf{H}^{(2)},\mathbf{M}^{(2)})$ are the solution to problem (1.14)–(1.16). Then, it follows from the boundedness of $\mathbf{M}^{(j)}$ and assumption (1.8) that

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\mathbf{M}^{(1)}(t)-\mathbf{M}^{(2)}(t)\|^2_{L^2(G)} \qquad (6.15)\\
&= -\int_G \gamma(\mathbf{M}^{(1)}(t)-\mathbf{M}^{(2)}(t))\cdot(\mathbf{M}^{(1)}(t)\wedge\mathbf{H}^{(1)}(t)-\mathbf{M}^{(2)}(t)\wedge\mathbf{H}^{(2)}(t))dx\\
&\quad+\int_G(\mathbf{M}^{(1)}(t)-\mathbf{M}^{(2)}(t))\cdot(\mathbf{a}(x,\mathbf{M}^{(1)}(t))-\mathbf{a}(x,\mathbf{M}^{(2)}(t)))dx\\
&\le -\int_G \gamma(\mathbf{M}^{(1)}(t)-\mathbf{M}^{(2)}(t))\cdot\mathbf{M}^{(1)}(t)\wedge(\mathbf{H}^{(1)}(t)-\mathbf{H}^{(2)}(t))dx\\
&\quad+C_1\|\mathbf{M}^{(1)}(t)-\mathbf{M}^{(2)}(t)\|^2_{L^2(G)}\\
&\le C_2\|\mathbf{M}^{(1)}(t)-\mathbf{M}^{(2)}(t)\|^2_{L^2(G)}+C_2\|\mathbf{H}^{(1)}(t)-\mathbf{H}^{(2)}(t)\|^2_{L^2(G)}.
\end{aligned}$$

Invoking Remark 1 and (1.14) one obtains

$$\mathbf{H}^{(1)}(t)-\mathbf{H}^{(2)}(t) = -(1-P_H)\left(\tilde{\mathbf{M}}^{(1)}(t)-\tilde{\mathbf{M}}^{(2)}(t)\right),$$

in particular

$$\|\mathbf{H}^{(1)}(t)-\mathbf{H}^{(2)}(t)\|^2_{L^2(\mathbb{R}^3)}\le C_3\|\mathbf{M}^{(1)}(t)-\mathbf{M}^{(2)}(t)\|^2_{L^2(G)}. \tag{6.16}$$

Finally, it follows from (6.15) and (6.16) that $\mathbf{M}^{(1)}=\mathbf{M}^{(2)}$ and also $\mathbf{H}^{(1)}=\mathbf{H}^{(2)}$ by (1.14) again. ∎

REFERENCES

[1] J. M. BALL, *Strongly continuous semigroups, weak solutions and the variation of constants formula*, Proc. Amer. Math. Soc., 63 (1977), pp. 370–373.

[2] F. BROWN, *Micromagnetics*, Wiley, New York, 1963.

[3] G. CARBOU AND P. FABRIE, *Time average in micromagnetism*, J. Differential Equations, 147 (1998), pp. 383–409.

[4] G. CARBOU, P. FABRIE, AND F. JOCHMANN, *A remark on the weak ω-limit set for micromagnetism equation*, Appl. Math. Lett., 15 (2002), pp. 95–99.

[5] P. DONNAT AND J. RAUCH, *Global solvability of the Maxwell Bloch equations from nonlinear optics*, Arch. Ration. Mech. Anal., 136 (1996), pp. 291–303.

[6] P. GERARD, *Microlocal defect measures*, Comm. Partial Differential Equations, 16 (1991), pp. 1761–1794.

[7] F. Jochmann, *The semistatic limit for Maxwell's equations in an exterior domain*, Comm. Partial Differential Equations, 23 (1998), pp. 2035–2076.
[8] F. Jochmann, *Long time asymptotics of solutions to the anharmonic oscillator model from nonlinear optics*, SIAM J. Math. Anal., 32 (2000), pp. 887–915.
[9] F. Jochmann, *Convergence to stationary states in the Maxwell Bloch system from nonlinear optics,* Quart. Appl. Math., 60 (2002), pp. 317–339.
[10] J. L. Joly, G. Metivier, and J. Rauch, *Global solvability of the anharmonic oscillator model from nonlinear optics*, SIAM J. Math. Anal., 27 (1996), pp. 905–913.
[11] J. L. Joly, G. Metivier, and J. Rauch, *Global solutions to Maxwell's equations in a ferromagnetic medium*, Ann. Henri Poincaré, 1 (2000), pp. 307–340.
[12] P. Joly, A. Komech, and O. Vacus, *On transitions to stationary states in a Maxwell–Landau–Lifschitz–Gilbert system*, SIAM J. Math. Anal., 31 (1999), pp. 346–374.
[13] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, Pergamon Press, New York, 1960.
[14] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[15] R. Picard, *An elementary proof for a compact embedding result in generalized electromagnetic theory*, Math. Z., 187 (1984), pp. 151–161.
[16] L. Tatar, *H-measures, a new approach for studying homogenization, oscillations and concentrations effect in partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 113 (1989), pp. 87–97.
[17] A. Visintin, *On Landau-Lifschitz equation for ferromagnetism*, Japan J. Appl. Math., 2 (1985), pp. 69–84.
[18] C. Weber, *A local compactness theorem for Maxwell's equations*, Math. Methods Appl. Sci., 2 (1980), pp. 12–25.
[19] N. Weck, *Maxwell's boundary value problem on Riemannian manifolds with nonsmooth boundaries*, J. Math. Anal. Appl., 46 (1974), pp. 410–437.
[20] H. Wynled, *Ferromagnetism,* in Encyclopedia of Physics, Vol. XVIII/2, Springer-Verlag, Berlin, 1966, pp. 25–36.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 341–358

THE 2-DIMENSIONAL RIEMANN PROBLEM FOR A 2×2 HYPERBOLIC CONSERVATION LAW I. ISOTROPIC MEDIA*

WOONJAE HWANG† AND W. BRENT LINDQUIST‡

Abstract. We construct the solutions for a two-dimensional (2-D) Riemann problem for a 2×2 hyperbolic nonlinear system based upon the Keyfitz–Kranzer–Isaacson–Temple model. The system is applicable to polymer flooding of an oil reservoir; the parameterization can be adjusted to model either isotropic or anisotropic media. For isotropic media, the solutions are obtained by two methods. The first method utilizes a transformation into a one-dimensional (1-D) Cauchy problem. Such a transformation requires conformity of the x- and y-directional fluxes in the system. The second method involves a 2-D constructive technique which can be used more generally for solving systems. For the isotropic media case, we explicitly construct solutions for the so-called single and four quadrant Riemann problems by both methods and demonstrate the equality of the solutions. This has relevance as a test for the 2-D solution method, as existence and uniqueness results for solutions of systems in 1-D are known, whereas no such results exist for systems in 2-D.

Key words. Riemann problems, hyperbolic systems, conservation law

AMS subject classifications. Primary, 35C05; Secondary, 35L65

PII. S0036141001396631

1. Introduction.

The existence and uniqueness theory for the scalar hyperbolic equation in multiple space dimensions is largely complete [4, 15, 26, 27]. The theory gives little insight into the form of the solutions which (e.g., [31, 24]) can possess interesting qualitative behavior. Since the solutions to multidimensional Riemann problems are also important for numerical computation, recent literature has concentrated on finding the solution to the general Riemann problem in two space dimensions. As no general theory exists for systems in multiple space dimensions, the two-dimensional (2-D) Riemann problem for systems must be computed on a case by case basis. Thus, our general interest is the achievement of a 2-D constructive technique which would be generally applicable for systems.

The study of the 2-D Riemann problem was initiated by Guckenheimer [10]. Further analyses of the 2-D scalar conservation law (1.1),

$$s_t + f(s)_x + h(s)_y = 0, \tag{1.1}$$

have appeared in [17, 18, 28, 31, 32, 33]. In a 2-D Riemann problem, the initial data is piecewise constant in wedge-shaped regions surrounding the origin. Wagner [28] constructed the solution for the four quadrant (90 degree regions) Riemann problem in the case of $f \equiv h$, where f, h are convex. Lindquist [17] showed that unique solutions to the arbitrary wedge problems are piecewise smooth when $f \equiv h$ and f has at most one inflection point. In a companion work [18], Lindquist outlined a systematic

*Received by the editors October 18, 2001; accepted for publication (in revised form) April 25, 2002; published electronically October 31, 2002. This work was supported by the Applied Mathematics Subprogram of the U.S. Department of Energy grant DE-FG02-90ER25084.
http://www.siam.org/journals/sima/34-2/39663.html
†Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609-2280 (woonjae@wpi.edu).
‡Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794-3600 (lindquis@ams.sunysb.edu).

method for construction of such solutions in terms of 2-D nonlinear waves. Zhang and Zheng [33] constructed the solution for the four quadrant Riemann problem under the extended condition $f_{ss} \neq 0$, $h_{ss} \neq 0$, and $\partial_s (f_{ss}/h_{ss}) \neq 0$. Chen, Li, and Tan [3] studied the dependence of the structure of the solution on the initial data values as well as wedge angles for the particular case of Riemann data arranged in three wedges.

A body of work has also developed for the important case of the Euler system of conservation laws modeling gas dynamics in two dimensions. A good summary is provided in the book by Chang (Zhang) and Hsiao [1]. Glimm et al. [6] presented a list of generic, steady (in some reference frame) waves (referred to as "nodes" by the authors) expected in 2-D Riemann problem solutions of the Euler equations. In [34], Zhang and Zheng presented conjectures on the classification and structure of the 2-D solutions to the four quadrant Riemann problem for the Euler equations of gas dynamics with a polytropic equation of state. In preparatory work for this conjecture, Tan and Zhang [22, 23] constructed analytic solutions for the simplified model

$$u_t + (u^2)_x + (uv)_y = 0, \tag{1.2}$$

$$v_t + (uv)_x + (v^2)_y = 0. \tag{1.3}$$

Yang and Zhang [29] verified the analytic solutions for (1.2), (1.3) numerically using the maximum-minimum bounds (MmB) preserving scheme. In [19], Schulz-Rinne presented a correction to the conjectured classification for the Euler equations. He showed that one of the solutions classified from the conjecture [34] is impossible. Schulz-Rinne, Collins, and Glaz [20] computed numerical Riemann problem solutions to the Euler equations in gas dynamics using the second order Godunov method and confirmed that, with the exception of one case, the conjectured solutions agree closely with the numerical results. Chang (Zhang), Chen, and Yang [2] performed numerical simulation for the Euler equations in gas dynamics with the MmB scheme to check the conjecture [34]. Lax and Liu [16] demonstrated that the numerical solution for the Euler equations obtained with their positive scheme are strikingly consistent with calculations by Schulz-Rinne, Collins, and Glaz [20]. Zhang and Zheng [35] obtained exact spiral solutions of the 2-D Euler equations. Zhang, Li, and Zhang [30] considered the 2-D Riemann problem for the pressure-gradient equations of the Euler system in the case of four quadrant initial data.

Other related work on Riemann problems has provided qualitative insights on 2-D wave interactions [5]; introduced a new type of nonlinear hyperbolic wave, a delta-shock wave, which is a Dirac delta function supported on a shock [24]; solved the 2-D Riemann problem for the transportation equations in the case of four quadrant initial data [21]; studied the solution of the 2-D Riemann problem of Hamilton Jacobi equations[7]; and examined Riemann problems in higher dimensions [8, 9].

In this paper and in part II [11], our interest is twofold.

The first is to develop the solutions to a class of 2-D Riemann problems for a 2×2 hyperbolic system based upon the Keyfitz–Kranzer–Isaacson–Temple model [14, 12, 25] which provides a model for polymer flood recovery in an oil reservoir with either an isotropic or anisotropic medium. This 2×2 system exhibits the distinctive feature that, regardless of the flux function, the second wave family is linearly degenerate, exhibiting only contact discontinuities and constant states. In this sense the system is the "simplest 2×2 system" beyond a scalar equation.

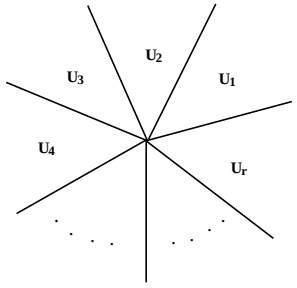


FIG. 1. *General 2-D Riemann problem initial data.*

Our second interest is the continued development of a 2-D Riemann problem solution constructive technique which has been generalized from that used in [18] for scalar equations. Recently, Zhang and Zhang [31] have classified the shock (S) and rarefaction (R) elementary waves into four types, S^+, S^-, R^+ and R^-, discussed the allowed interactions amongst them, and generalized the characteristic based solution construction method. The constructive method is based upon the existence of base points and base curves in the space of variables $\xi = x/t, \eta = y/t$ which govern the location of shocks, rarefactions and contact discontinuities in the self-similar growth of the 2-D Riemann problem solution. In part II of this paper [11] we show the existence of both flux function dependent and flux function independent relationships among these points and curves. In this paper, our interest in the constructive technique is to show, by example, that the technique is capable of providing correct, unique solutions (when they exist) to a 2×2 system Riemann problem.

For the isotropic media parameterization of our model, solutions can be obtained by two methods. The first method utilizes a transformation into a one-dimensional (1-D) Cauchy problem. (Such a transformation requires conformity of the x- and y-directional fluxes in the system.) The second method is via the 2-D constructive technique. For the isotropic media case, we construct solutions by both methods and demonstrate equality. This provides a test for the 2-D solution method, as existence and uniqueness results for solutions of systems in one dimension are known, whereas no such results exist for systems in two dimensions.

The Riemann problem in two spatial dimensions for a system of n conservation laws is the problem

$$U_t + F(U)_x + H(U)_y = 0, \tag{1.4}$$

with solution and flux vectors $U = (u_1, u_2, \ldots, u_n)$, $F = (f_1(U), f_2(U), \ldots, f_n(U))$, $H = (h_1(U), h_2(U), \ldots, h_n(U))$, and initial data that is piecewise constant on a finite number r of wedges centered on the origin $x = 0$, $y = 0$ as shown in Figure 1; i.e.,

$$U(0, x, y) = U_i, \qquad i = 1, \ldots, r,$$

where, for each i, x and y lie in the wedge between the half lines determined by the parametric equations

$$\left[\begin{array}{rcl} x_i & = & \tau \cos(\theta_i), \\ x_{i+1} & = & \tau \cos(\theta_{i+1}), \end{array} \quad \begin{array}{rcl} y_i & = & \tau \sin(\theta_i) \\ y_{i+1} & = & \tau \sin(\theta_{i+1}) \end{array} \right], \qquad 0 \le \tau < \infty, \quad r+1 \equiv 1$$

for r ordered, counterclockwise positive angles $0 \le \theta_1 < \theta_2 < \cdots < \theta_r < 2\pi$. Of particular interest is the four wedge problem with wedges corresponding to the four

quadrants ($\theta_1 = 0, \theta_2 = \pi/2, \theta_3 = \pi, \theta_4 = 3\pi/2$) of the spatial plane, since such initial data is pertinent to numerical finite difference schemes. We also consider the "single quadrant" Riemann problem having $\theta_1 = \pi/2, \theta_2 = \pi$. (The choice of the quadrant is discussed in section 3.1.)

In much of the work on systems [16, 19, 20, 21, 22, 23, 24, 29, 30, 34, 35], the initial Riemann data is restricted so that the presence of a single wave is always guaranteed to develop from each initial discontinuity. This has some physical justification, in that the majority of physical observations involve the study of a single propagating wave type. It is, nonetheless, a restrictive assumption when considering wave interactions, which in general, for an $n \times n$ system, would produce n postinteraction waves. For this reason we make no such restriction on the initial Riemann data in the development of our solutions.

In section 2, we present the specifics of our 2×2 system. In section 3, we consider 2-D Riemann problems for (1.4) for the isotropic media parameterization of our model: $H(U) = \alpha F(U)$, where α is a constant. Both single quadrant and four quadrant initial Riemann data are considered. Solutions are developed by both methods and compared.

2. The model. The Keyfitz–Kranzer–Isaacson–Temple 2×2 system of conservation laws in one space dimension is

$$s_t + f(s,c)_x = 0, \tag{2.1}$$

$$(cs)_t + (cf(s,c))_x = 0, \tag{2.2}$$

where the physical state variables $U = (s, c)$ are

$$\begin{aligned} s &= \text{water saturation}, & 0 \le s \le 1, \\ c &= \text{concentration of polymer}, & 0 \le c \le 1. \end{aligned}$$

This system models the polymer flooding of an oil reservoir by generalizing the Buckley–Leverett equation which models a two phase water-flood process. In the polymer flood, a (generally small) amount of polymer is added to the water to increase the sweep efficiency of oil production. The model assumes the polymer is completely miscible in the water phase and undergoes no mass transfer into the oil phase. See [12] for details of the model.

Regardless of the form of f, this model contains two families of waves, a c-family consisting of a contact discontinuity, and an s-family identical to the scalar family (2.1) with c held a constant. The functional forms for $f(\cdot,\cdot)$ commonly used with this model in enhanced oil recovery [12, 13] and elasticity [14] studies contain a single inflection point, and the s-family thus consists of compound waves composed of a Lax shock and a rarefaction fan.

In order to partially eliminate the complexity introduced by the compound waves in the s-family, we simplify to a convex function f. A representative example for f is

$$f(s,c) = s^2[1 + A(1-c)(1-s)], \qquad 0 < A < 1/2, \qquad 0 \le s, c \le 1, \tag{2.3}$$

having the form shown in Figure 2.

In analyzing the solutions to (2.1), (2.2) it is convenient to introduce the change of variables

$$s,\ c \to s,\ b \equiv sc,$$

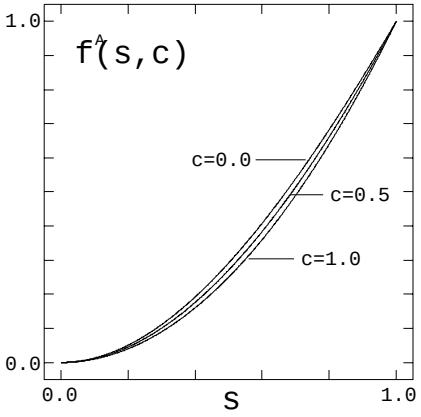


FIG. 2. *The flux function for the model.*

and introduce

$$g \equiv f(s,c)/s.$$

For smooth solutions to (2.1), (2.2) the eigenvalues, right eigenvectors, and Riemann invariants for the two families of waves are

$$\begin{aligned} \lambda^s &= f_s, & \lambda^c &= g, \\ r^s &= (s,b), & r^c &= (-g_b, g_s), \\ W^s &= c, & W^c &= g, \end{aligned}$$

independent of the form of f [12]. The Rankine–Hugoniot relations, also independent of the form of f, are

$$\begin{aligned} c_l &= c_r, & g_l &= g_r, \\ \sigma_s &= \frac{[f]}{[s]}, & \sigma_c &= g_l = g_r. \end{aligned}$$

For convex functions such as (2.3), the system remains strictly hyperbolic; the eigenvalues of the two families coinciding only on the axis $s = 0$. For a strictly hyperbolic system, Figure 3 shows the Hugoniot/rarefaction solution curves for the four general cases of a 1-D Riemann problem $U_L \to U_{R_i}$. U_{M_j} denotes the intermediate state in each solution.

We extend the model (2.1), (2.2) to two space dimensions in a manner that has physical relevance. For spatially isotropic media, the extension is

$$H(U) = \alpha F(U), \qquad F(U) \equiv (f, cf), \tag{2.4}$$

with α = constant. For anisotropic media, having principle axes aligned in the x and y directions, $F(U)$ and $H(U)$ would most likely differ in the value of one or more physical parameters, for example, in the value of A in (2.3). Our model for anisotropic media is then

$$s_t + f^A(s,c)_x + f^B(s,c)_y = 0, \tag{2.5}$$

$$(cs)_t + (cf^A(s,c))_x + (cf^B(s,c))_y = 0, \tag{2.6}$$

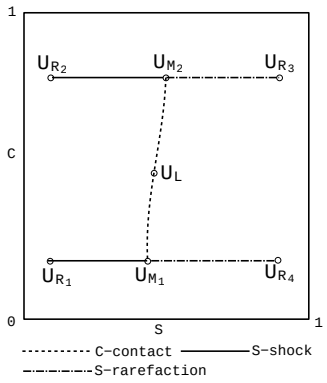


FIG. 3. *The Hugoniot and rarefaction curves for the four general solution cases to the* 1*-D Riemann problem in the model.*

where

$$
\begin{aligned}
f^A(s,c) &= s^2[1 + A(1-c)(1-s)], \qquad \text{with } 0 < A < 1/2, \\
f^B(s,c) &= s^2[1 + B(1-c)(1-s)], \qquad \text{with } 0 < B < 1/2.
\end{aligned} \tag{2.7}
$$

s and c have the same range of values as in the 1-D model.

There are several reasons for considering this model. One is physical, its relevance for flow in porous media. The remainder are mathematical. Note that system (2.5), (2.6) can be written

$$s_t + (sg^A(s,b))_x + (sg^B(s,b))_y = 0, \tag{2.8}$$

$$b_t + (bg^A(s,b))_x + (bg^B(s,b))_y = 0, \tag{2.9}$$

with $g^\alpha(s,b) \equiv f^\alpha(s,b)/s,\ \alpha = A, B$. Any system of this form remains hyperbolic in any spatial direction regardless of the form of the (real-valued) functions g^A and g^B! The model can be viewed as the simplest extension from a scalar equation to a hyperbolic system in the sense that the second family of waves introduced (the c-wave family) is linearly degenerate, producing only contact discontinuity waves. The restrictions (2.7) on the flux functions guarantee $f^A_{ss} \neq 0$, $f^B_{ss} \neq 0$ and $\frac{\partial}{\partial s}(f^A_{ss}/f^B_{ss}) \neq 0$ for any $s \in (0,1]$, $c \in [0,1]$. This in turn guarantees [33] that the s-wave family is genuinely nonlinear in all spatial directions except one (along which it becomes linear).

As we shall see, a result of the linear degeneracy of the second wave family is a division of the solution space t, x, y into regions where the solution is governed by a scalar equation, the boundary between these regions being the contact discontinuity. This will ultimately guarantee the computation of globally unique solutions.

As the 2-D Riemann problem is self-similar, the solution needs to be constructed only in a single t = const plane; the usual choice is $t = 1$. The construction method developed in [18, 31] assembles the global solution in this plane by "joining together" elementary waves (shocks, rarefactions, their composite waves, and contacts). For scalar equations in two dimensions, global uniqueness of the solution results implicitly by ensuring the elementary waves obey local entropy conditions due to Vol'pert [27] or Kruzkov [15]. For systems, no such general local entropy

conditions exist and global uniqueness becomes an issue. In particular, no general uniqueness theory exists that is applicable to the system under consideration here.

The 2-D Riemann problem for the case of isotropic media (2.4) is the $B \to A$ limiting case of anisotropic media. We deal with the isotropic case separately in this paper since it provides information both as to the form and uniqueness of the global solutions for the more general anisotropic problem.

The isotropic case is amenable to two solution methods. Under rotation by $\theta = \tan^{-1}\alpha$, the 2-D problem converts into uncoupled, 1-D Cauchy problems

$$U_t + \sec(\theta)F(U)_\xi = 0, \qquad (x, y) \to (\xi, \eta), \tag{2.10}$$

whose initial data can be described as interacting Riemann problems [17, 18]. Existence and uniqueness for this system in one dimension is known [12]. Thus comparison of the rotated 1-D and direct 2-D construction methods is a check that the direct 2-D construction method is producing the correct unique global solution. In fact, as we shall see for this system, all 2-D solutions consist of two separate solution regions whose common boundary is a contact discontinuity. In each of these two regions the variable c is constant in value; in each region the solution reduces to that of a scalar Riemann problem, to which the existence and uniqueness conditions of Vol'pert and Kruzkov can be applied. The comparison with the rotated 1-D solution thus provides verification that this spatial decomposition into two regions separated by the contact discontinuity is unique. This uniqueness is exploited in part II of this paper [11].

3. Isotropic media. We consider the system

$$s_t + f(s, c)_x + \alpha f(s, c)_y = 0, \tag{3.1}$$

$$(cs)_t + (cf(s, c))_x + \alpha(cf(s, c))_y = 0 \tag{3.2}$$

with the form of f consistent with the behavior of example (2.3). Without loss of generality, we take $\alpha = 1$. Under the transformation

$$\xi = (x + y)/2, \qquad \eta = (y - x)/2$$

(a rotation of $\pi/4 = \tan^{-1}(1)$ combined with a dilation by a factor of $\sqrt{2}$), the system (3.1), (3.2) becomes

$$s_t + f(s, c)_\xi = 0, \tag{3.3}$$

$$(cs)_t + (cf(s, c))_\xi = 0. \tag{3.4}$$

From (3.3), (3.4) we see that the solution can be obtained in each $(\xi, \eta = \text{ const}, t)$ plane independent of other η. In particular, this implies the solutions obtained in the $\eta < 0$ half-space can be obtained independently of the solutions in the $\eta > 0$ half-space. The catalog of solutions in both half-spaces is the same; for any particular 2-D Riemann problem, the solutions in the two half-spaces join in a consistent manner along the $\eta = 0$ axis [17, 18]. We therefore restrict our discussion to the half-space $\eta > 0$ $(y > x)$.

Solution topology depends on initial data values in the Riemann problem wedges. We present the complete catalog of solution topologies for the single and four quadrant Riemann problems in section 3.1 and section 3.2, respectively.

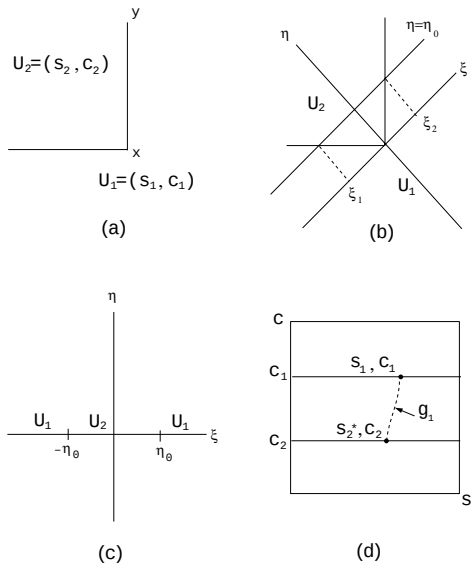


FIG. 4. *The single quadrant Riemann problem:* (a) *the initial Riemann data;* (b) *the wedge geometry;* (c) *the initial* 1*-dim Cauchy data along a line* $\eta = \eta_0$*;* (d) *the solution wave curves for the transition* $(s_1, c_1) \to (\cdot, c_2)$ *producing the intermediate state* (s_{2^*}, c_2).

3.1. The single quadrant Riemann problem. For the single quadrant problem the wedge angles are $\theta_1 = \pi/2$ and $\theta_2 = \pi$. The initial Riemann problem data is shown in Figure 4(a) and the rotated (ξ, η) coordinate system is sketched in Figure 4(b). Figure 4(c) shows the initial 1-D Cauchy data along a line $\eta = \eta_0$. Numbering the four quadrants clockwise from the negative y-axis (see Figure 9(a)) we note that of the four possible single quadrant choices, choosing 2 or 4 will produce equivalent solutions; choosing 1 or 3 leads to trivial solutions as either quadrant choice results in two initial data discontinuities, one above and one below the $\eta = 0$ plane. In this case waves emanating from the discontinuity in the $\eta > 0$ plane propagate independently of those emanating from the discontinuity in the $\eta < 0$ plane.

Avoiding the choice $c_1 = c_2$ which reverts to the scalar problem, there are four possible classes of initial data U_1 and U_2 for the single quadrant problem:

$$\begin{array}{ll} \text{case 1} & c_1 > c_2, \quad s_{2^*} \geq s_2; \\ \text{case 2} & c_1 > c_2, \quad s_{2^*} < s_2; \\ \text{case 3} & c_1 < c_2, \quad s_{2^*} \geq s_2; \\ \text{case 4} & c_1 < c_2, \quad s_{2^*} < s_2, \end{array}$$

where the intermediate state $U_{2^*} = (s_{2^*}, c_2)$ is defined in Figure 4(d). Only the first two cases lead to topologically distinct solutions; cases 1 and 3 have the same structure as cases 2 and 4. We label the two distinct topologies by their structure "at spatial infinity" (i.e., $R \to \infty, R^2 \equiv x^2 + y^2$). Thus case 1 (and case 3) topology is characterized by a contact and a shock wave propagating in the y-direction, a contact and rarefaction wave propagating in the x-direction. This topology is labeled CSCR. Case 2 (and case 4) topology is reversed and is labeled CRCS.

3.1.1. The single quadrant CSCR solution. We first present the solution obtained by rotating into a 1-D Cauchy problem. We will be brief in our discus-

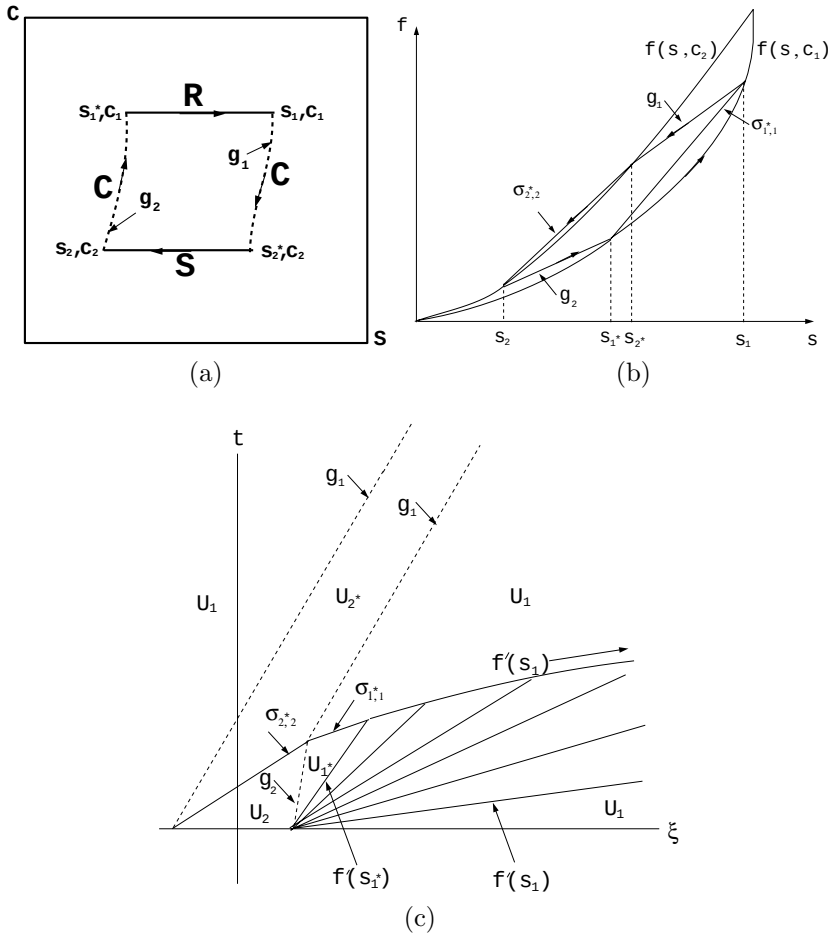


FIG. 5. *CSCR solution of the single quadrant Riemann problem:* (a) *wave curves in phase space for the Riemann problems at* $\xi = \pm\eta_0$ *(C, S, and R represent contact discontinuity, shock, and rarefaction wave, respectively);* (b) *the solution construction on the flux functions for the Riemann problems at* $\xi = \pm\eta_0$ *and for the shock separating the states* U_{1*} *and* U_1*;* (c) *the solution in the* t, ξ, η_0 *plane.*

sion of the solution as it involves well-known general 1-D Cauchy problem solution construction methods based upon the particular Riemann problem solution for this model [12].

On the line $\eta = \eta_0$, the initial data discontinuities are positioned at $\xi = \pm\eta_0$. The $U_1 \to U_2$ transition at $\xi = -\eta_0$ results in a contact (speed g_1) and shock (speed $\sigma_{2^*,2}$) wave and an intermediate state U_{2^*}. The $U_2 \to U_1$ transition at $\xi = \eta_0$ results in a contact (speed g_2) and rarefaction (characteristic speeds between $f'(s_{1^*})$ and $f'(s_1)$) fan and an intermediate state U_{1^*}. Figure 5(a) shows the phase space solution for these two separate Riemann problems. Details on solving for the states $U_{1^*} = (s_{1^*}, c_1)$ and $U_{2^*} = (s_{2^*}, c_2)$ are given in the appendix.

The shock emanating from $\xi = -\eta_0$ and the contact from $\xi = \eta_0$ interact at a later time, resulting in a Riemann problem between the states U_{2^*} and U_{1^*}. This Riemann problem produces a contact (speed g_1) and a shock $\sigma_{1^*,1}$. Figure 5(b) presents the

solution construction on the flux functions for the Riemann problems at $\xi = \pm\eta_0$ for this contact-shock interaction.

The shock of speed $\sigma_{1^*,1}$ interacts with the rarefaction fan. As a result the shock speed increases, approaching an asymptotic speed of $f'(s_1)$. Figure 5(c) shows the entire solution in the t, ξ, η_0 plane. The solution in any other plane t, ξ, $\eta > 0$ is self-similar to this.

Basic principles for the direct 2-D Riemann problem solution construction method are outlined in [18, 31]. (The terminologies are slightly different.) The 2-D solution method is discussed fully in part II of this paper [11] for application to the anisotropic medium model. This discussion carries over to the isotropic model by setting $B = A$. With a view to introducing necessary notation, the relevant principles in the construction method, specialized to the isotropic medium model, are briefly stated here.

• The construction is performed in the $t = 1, x, y$ plane. For brevity, we henceforth suppress the $t = 1$ label for points in this plane.

• If p is a point (in this x, y plane) on a c-family contact discontinuity separating the two states (s_l, c_l) and (s_r, c_r), the tangent line to the contact discontinuity at p passes through the two *contact base points* g_l and g_r having coordinate values

$$g_l = (g(s_l, c_l), g(s_l, c_l)) \quad \text{and} \quad g_r = (g(s_r, c_r), g(s_r, c_r)). \tag{3.5}$$

From the contact discontinuity Rankine–Hugoniot conditions, these two points are identical, $g_l = g_r$.

• If p is a point on an s-family shock separating the two states (s_l, c) and (s_r, c), the tangent line to the shock at p passes through the *shock base point* σ_{lr} having coordinate values

$$\sigma_{lr} = (\sigma(s_l, s_r), \sigma(s_l, s_r)), \tag{3.6}$$

where

$$\sigma(s_l, s_r) \equiv \frac{f(s_l) - f(s_r)}{s_l - s_r}. \tag{3.7}$$

• s-family shocks develop in regions of constant c, i.e., in regions where the problem reduces to being scalar. Thus Kruzkov's entropy conditions are applicable, these are shown in [11] to reduce to familiar Lax shock conditions and become easily verifiable angle restrictions in the case of the isotropic model.

• An s-family level curve (i.e., characteristic in an s-family rarefaction fan) having state value (s_p, c_p) is a straight line segment whose tangent passes through the *characteristic base point* $f'(p)$ having coordinate values

$$f'(p) = (f_s(s_p, c_p), f_s(s_p, c_p)), \tag{3.8}$$

where the subscript represent the differentiation with respect to s.

The 2-D construction of solutions for the case of an isotropic medium simplifies since all base points lie along the line $y = x$ in the $t = 1$ plane. (This is not true in the anisotropic medium case.)

Figure 6 displays the solution obtained by the 2-D construction method in the $t = 1, x, y$ plane. The solution for any other $t > 0$ is self-similar to this. Constant states and relevant base points on the line $y = x$ are labeled. As stated, our interest in this paper is to verify that the 2-D construction method produces the correct globally unique solution by comparing against the rotated 1-D solutions. Thus

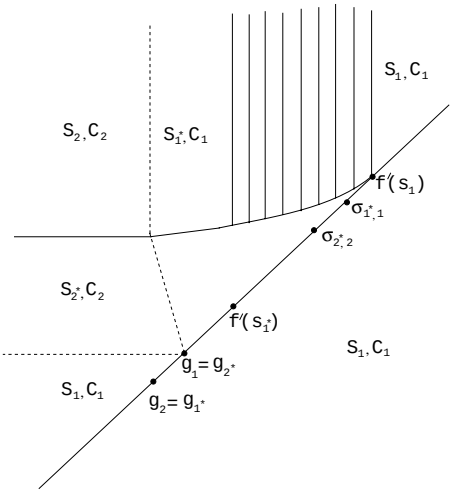


FIG. 6. *Direct 2-D construction of the CSCR solution of the single quadrant Riemann problem in the* $t = 1, x, y$ *plane.*

it is necessary to compare the solution in Figure 5(c) (which is self-similar in η) to the solution in Figure 6 (which is self-similar in t). While a mapping between these two solutions can be formally derived [17], examination of qualitative features of this map are enough to determine that the two solutions are indeed identical. First, a cut through the solution in Figure 6 along the line $y = x + 2\eta_0$ should be identical to a $t = 1$ cut through the solution in Figure 5. A cut through the solution in Figure 6 along any line $y = x + 2\eta$ for $\eta < \eta_0$ should correspond to a constant t cut through the solution in Figure 5 for some $t > 1$. As the value of $\eta \to 0$, the corresponding value for $t \to \infty$. Similarly, a cut through the solution in Figure 6 along any line $y = x + 2\eta$ for $\eta > \eta_0$ should correspond to a constant t cut through the solution in Figure 5 for some $t < 1$. As the value of $\eta \to \infty$, the corresponding value for $t \to 0$. Comparison of such qualitative features between the two figures easily verifies the identical nature of the two solutions.

3.1.2. The single quadrant CRCS solution. Figure 7(a)–(c) presents the analogous 1-D rotated solution construction for the CRCS topology solution. Figure 8 presents the direct 2-D solution construction. Again, qualitative comparison reveals that the direct construction is computing the globally unique solution.

We draw attention to an important characteristic feature of the solutions in Figures 6 and 8. The continuous contact discontinuity wave divides the solution into two regions, each region having constant concentration value c. In each such region, the solution is governed by a scalar equation in two dimensions. *Existence and uniqueness of the construction of the solution in these regions should be governed by multidimensional scalar theory, i.e., by the work of Vol'pert and Kruzkov, and hence the direct construction method employed, which is based upon the Rankine–Hugoniot and entropy conditions of this work, should produce a unique solution in each region.*

It is also interesting to note the physical diffraction present in the solution in Figure 8. The contact discontinuity separates regions of different polymer concentration. The polymer concentration is acting like an index of refraction. Rarefaction waves emerging from one region into the other are diffracted according to a "Snell's law."

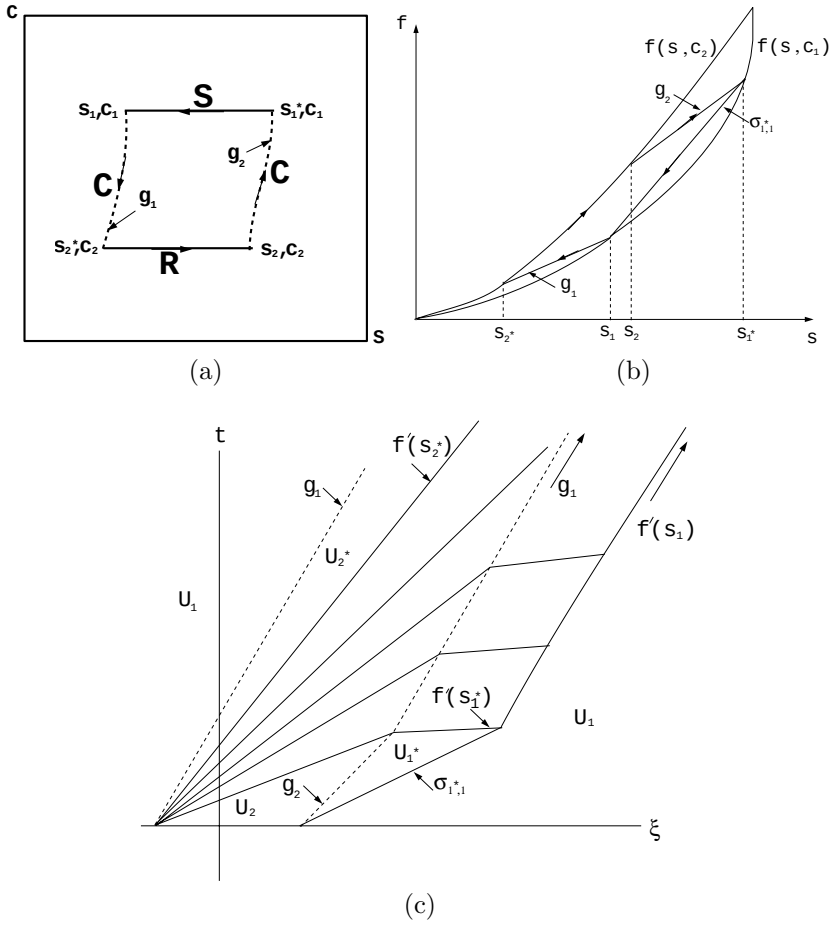


FIG. 7. *CRCS solution of the single quadrant Riemann problem:* (a) *wave curves in phase space for the Riemann problems at* $\xi = \pm\eta_0$ *(C, S, and R represent contact discontinuity, shock, and rarefaction wave, respectively);* (b) *the solution construction on the flux functions for the Riemann problems at* $\xi = \pm\eta_0$ *and for the shock separating the states* U_{1*} *and* U_1*;* (c) *the solution in the* t, ξ, η_0 *plane.*

The situation is dynamic, however, in that the interaction of the rarefaction waves with the contact also dynamically determines not only the strength of the diffraction but the boundary position at which the diffraction occurs.

3.2. The four quadrant Riemann problem. We now turn to the problem of initial data specified in the four quadrants (Figure 9(a)). We will present the solution in the $t = 1, x, y$ plane as computed by the direct 2-D solution method. For brevity, we do not show the 1-D rotated solutions; they are identical to those constructed directly in two dimensions.

As noted previously, we need only catalog the solution forms in the half-space $y \geq x$ ($\eta \geq 0$) as this produces the same catalog of solutions as in the half-space $y \leq x$. Thus we need not consider data in quadrant 4. Given initial data $U_1 = (s_1, c_1)$, $U_2 = (s_2, c_2)$, $U_3 = (s_3, c_3)$, we consider only the cases in which c_1, c_2, and c_3 all differ. There are 36 possible orderings of the initial data U_1, U_2, U_3, but only four topologically distinct solutions for the isotropic model. The four distinct cases, again

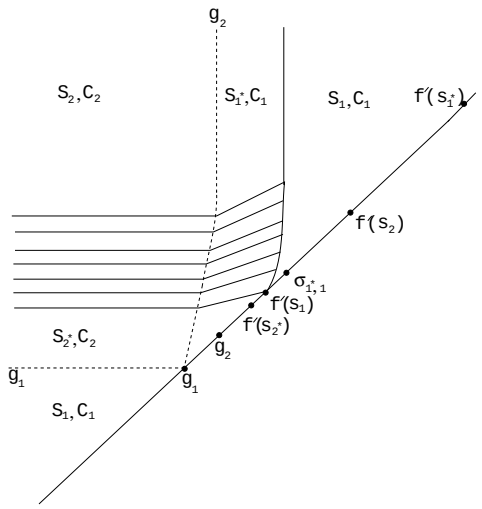


FIG. 8. *Direct 2-D construction of the CRCS solution of the single quadrant Riemann problem in the $t = 1, x, y$ plane.*

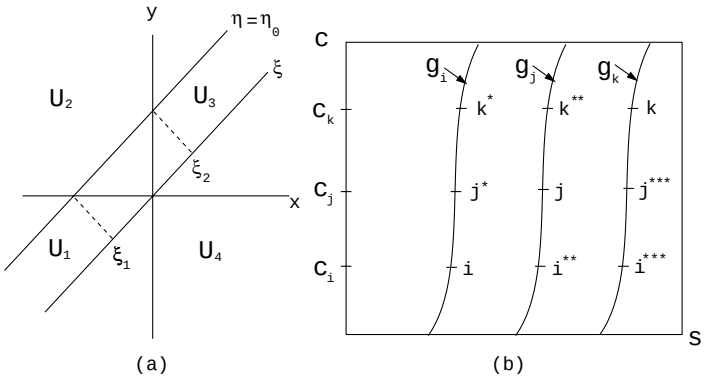


FIG. 9. *The four quadrant Riemann problem:* (a) *the initial Riemann data;* (b) *notation used for states relative to an initial state U_j. The state U_k is a final state having $s_k > s_j$, $c_k > c_j$. The state U_i is a final state having $s_i < s_j$, $c_i < c_j$. All other states are intermediate.*

labeled by the waves which appear "at spatial infinity," are

$$\begin{array}{lllll}
\text{case CRCR} & c_1 < c_2 < c_3, & s_1 \le s_{1^{**}}, & s_{3^{**}} \le s_3; \\
\text{case CSCS} & c_3 < c_2 < c_1, & s_3 \le s_{3^{**}}, & s_{1^{**}} \le s_1; \\
\text{case CSCR} & c_2 < c_1 \le c_3, & s_2 \le s_{2^{**}}, & s_{3^{**}} \le s_3; \\
\text{case CRCS} & c_1 \le c_3 < c_2, & s_1 \le s_{1^{**}}, & s_{2^{**}} \le s_2,
\end{array}$$

where, for $c_i < c_j < c_k$, the states $U_{i^{**}}$, U_j, $U_{k^{**}}$ lie on the c-family Hugoniot locus passing through state U_j as shown in Figure 9(b).

The phase space portrait of the Hugoniot and rarefaction curves for case CRCR are shown in Figure 10(a); relevant wave speed computations are shown in Figure 10(b). The solution in the $t = 1$, x, y plane, showing constant states and relevant base points, is given in Figure 10(c). The intermediate state saturation values s_{2^*}, s_{3^*}, and $s_{3^{**}}$ are found using the same procedure given in the appendix.

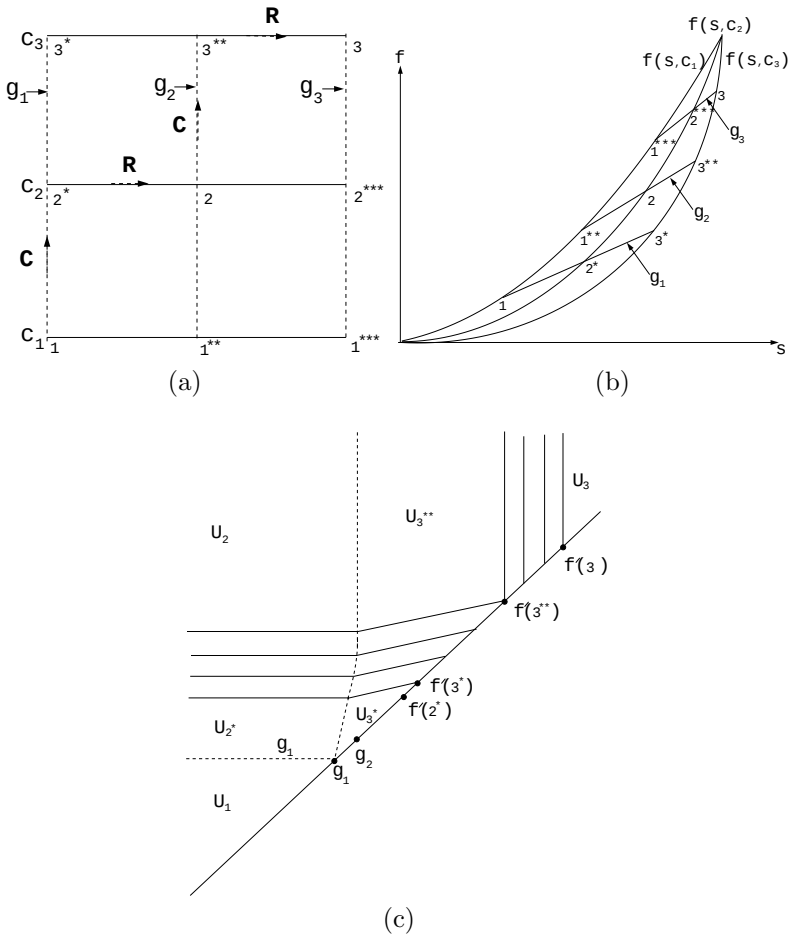


FIG. 10. *CRCR solution of the four quadrant Riemann problem:* (a) *wave curves in phase space and* (b) *solution construction on the flux functions for the relevant Riemann problems encountered in the solution; (C, S, and R represent contact discontinuity, shock, and rarefaction wave, respectively);* (c) *the solution in the* $t = 1, x, y$ *plane.*

Similar plots for the CSCS, CSCR, and CRCS cases are given, respectively, in Figures 11, 12, and 13.

If the $y < x$ solution half of the solution is included, then in the general case where c_1, c_2, c_3, and c_4 are all different in value, the solution will consist of four distinct regions in each of which c remains at the initial concentration value. These regions are separated by a piecewise smooth contact discontinuity boundary. In each region the solution is governed by a scalar equation, and the solution constructed by the direct 2-D method utilizing Rankine–Hugoniot and entropy conditions due to Kruzkov will be unique.

4. Discussion. We have verified that the 2-D Riemann problem construction method outlined in [11] produces the globally unique solution to the isotropic medium model (3.1), (3.2) for single and four quadrant initial data.

From the phase space constructions we note that, in this model, there is no method for introducing values of concentration that are not present in the initial

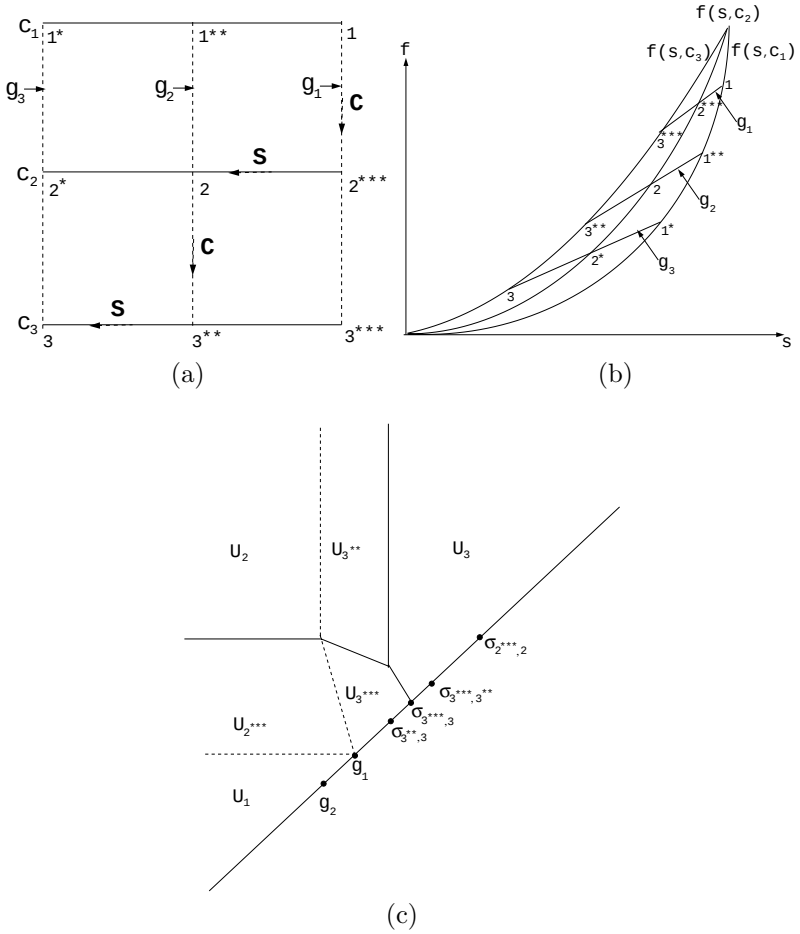


FIG. 11. *CSCS solution of the four quadrant Riemann problem:* (a) *wave curves in phase space and* (b) *solution construction on the flux functions for the relevant Riemann problems encountered in the solution; (C, S, and R represent contact discontinuity, shock, and rarefaction wave, respectively);* (c) *the solution in the* $t = 1, x, y$ *plane.*

data; similarly, there is no mechanism for eliminating a concentration value present in the initial data. This observation would hold for a Riemann problem consisting of a general (finite) number of wedges. Thus the general wedge 2-D Riemann problem solution must consist of r different regions of constant concentration, where r is the number of different concentrations in the initial data. These r regions will be joined by piecewise smooth contact discontinuities. The solution within each region will be governed by a scalar equation, and the solution constructed by the direct 2-D method utilizing Rankine–Hugoniot and entropy conditions due to Kruzkov will be unique.

5. Appendix. Computation of s_{1^*}. The saturation value for the intermediate state U_{1^*} in the single quadrant CSCR topology is computed as follows. For given values s_2, c_1 and c_2, s_{1^*} can be computed from the Rankine–Hugoniot relation

$$g^A(s_{1^*}, c_1) = g^A(s_2, c_2). \tag{A.1}$$

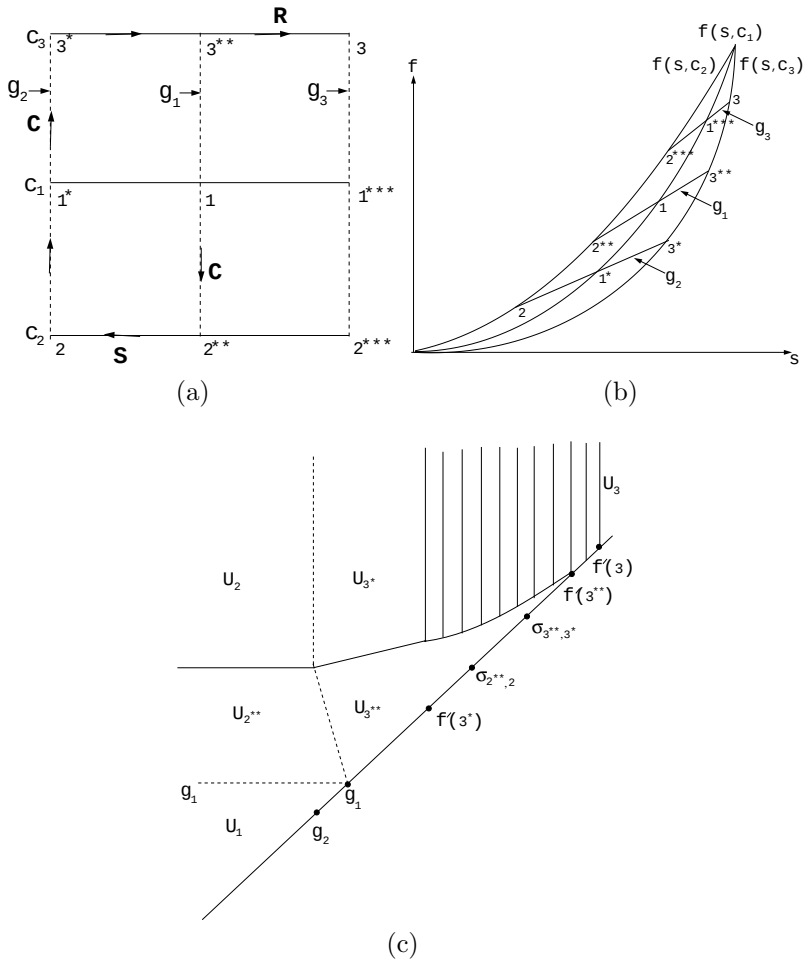


FIG. 12. *CSCR solution of the four quadrant Riemann problem:* (a) *wave curves in phase space and* (b) *solution construction on the flux functions for the relevant Riemann problems encountered in the solution; (C, S, and R represent contact discontinuity, shock, and rarefaction wave, respectively);* (c) *the solution in the* $t = 1, x, y$ *plane.*

Let $g_2^A \equiv g^A(s_2, c_2)$. Solving (A.1) for s_{1^*} produces the quadratic

$$A_1(s_{1^*})^2 - (1 + A_1)s_{1^*} + g_2^A = 0, \tag{A.2}$$

where $A_1 \equiv A(1 - c_1)$. The quadratic has two solutions s_+, s_- where $s_+ > s_-$. As $s_+ > 1$ and $s_- \in (0, 1)$, the appropriate value for s_{1^*} is s_-.

Computation of s_{2^*}. Computation of s_{2^*} proceeds as for s_{1^*} using the relation

$$g^A(s_{2^*}, c_2) = g^A(s_1, c_1). \tag{A.3}$$

Solving for s_{2^*} again results in a quadratic with the appropriate solution value being the unique root lying in the range $(0, 1)$.

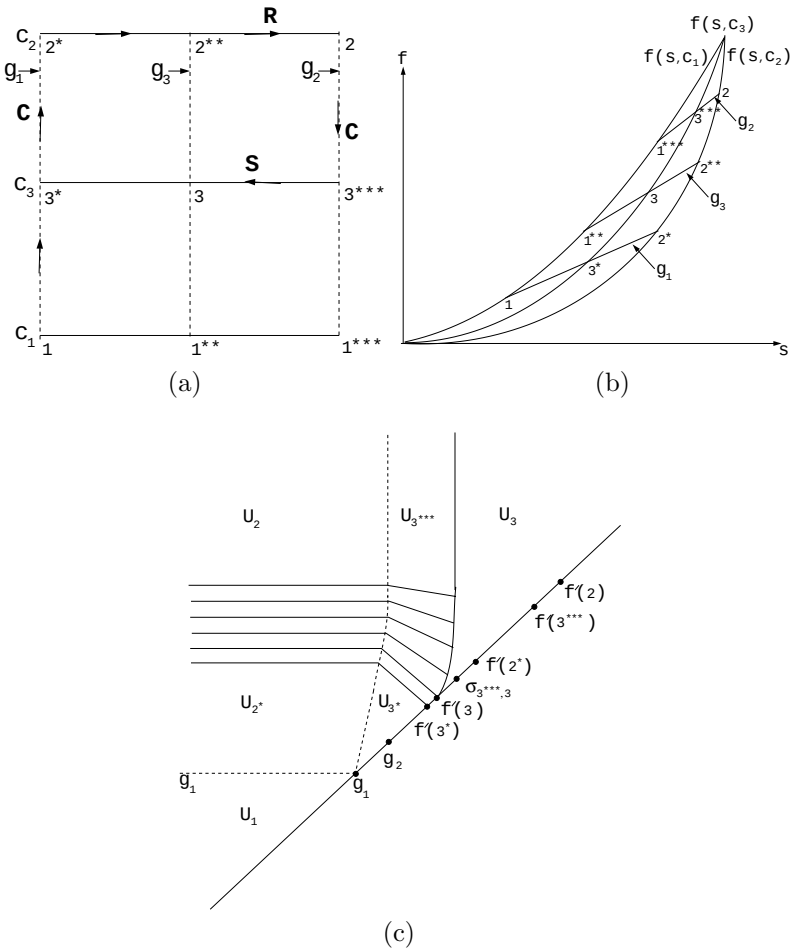


FIG. 13. *CRCS solution of the four quadrant Riemann problem:* (a) *wave curves in phase space and* (b) *solution construction on the flux functions for the relevant Riemann problems encountered in the solution; (C, S, and R represent contact discontinuity, shock, and rarefaction wave, respectively);* (c) *the solution in the* $t = 1, x, y$ *plane.*

REFERENCES

[1] T. CHANG AND L. HSIAO, *The Riemann Problem and Interaction of Waves in Gas Dynamics*, Pitman Monogr. Surveys Pure Appl. Math. 41, Longman Scientific and Technical, Harlow, UK, 1989.

[2] T. CHANG, G. CHEN, AND S. YANG, *On the Riemann problem for* 2*-D compressible Euler equations* I. *Interaction of shocks and rarefaction waves*, Discrete Contin. Dynam. Systems, 1 (1995), pp. 555–584.

[3] G. CHEN, D. LI, AND D. TAN, *Structure of Riemann solutions for* 2*-dimensional scalar conservation laws*, J. Differential Equations, 127 (1996), pp. 124–147.

[4] E. CONWAY AND J. SMOLLER, *Global solutions of the Cauchy problem for quasi-linear first order equations in several space variables*, J. Comm. Pure Appl. Math., 19 (1966), pp. 95–105.

[5] J. GLIMM, *The interaction of nonlinear hyperbolic waves*, Comm. Pure Appl. Math., 41 (1988), pp. 569–590.

[6] J. GLIMM, C. KLINGENBERG, O. MCBRYAN, B.PLOHR, D. SHARP, AND S. YANIV, *Front tracking and two-dimensional Riemann problems*, Adv. Appl. Math., 6 (1985), pp. 259–290.

[7] J. GLIMM, H.C. KRANZER, D. TAN, AND F.M. TANGERMAN, *Wave fronts for Hamilton-Jacobi equations: The general theory*, Comm. Math. Phys., 187 (1997), pp. 647–677.

[8] J. Glimm and D. Sharp, *An S matrix theory for classical nonlinear physics*, Found. Phys., 16 (1986), pp. 125–141.
[9] J. Glimm and D. Sharp, *Elementary Waves for Hyperbolic Equations in Higher Dimensions: An Example from Petroleum Reservoir Modeling*, Contemp. Math. 60, AMS, Providence, RI, 1987, pp. 35–41.
[10] J. Guckenheimer, *Shocks and rarefactions in two space dimensions*, Arch. Ration. Mech. Anal., 59 (1975), pp. 281–291.
[11] W. Hwang and W. B. Lindquist, *The* 2*-dimensional Riemann problem for a* 2×2 *hyperbolic conservation law* II. *Anisotropic media*, SIAM J. Math. Anal., 34 (2002), pp. 359–384.
[12] E. Isaacson, *Global Solution of a Riemann Problem for a Non-Strictly Hyperbolic System of Conservation Laws Arising in Enhanced Oil Recovery*, preprint, The Rockefeller University, New York, 1980.
[13] T. Johansen and R. Winther, *The solution of the Riemann problem for a hyperbolic system of conservation laws modeling polymer flooding*, SIAM J. Math. Anal., 19 (1988), pp. 541–566.
[14] B. Keyfitz and H. Kranzer, *A system of non-strictly hyperbolic conservation laws arising in elasticity theory*, Arch. Rational Mech. Anal., 72 (1980), pp. 219–241.
[15] S. N. Kruzkov, *First order quasilinear equations in several independent variables*, J. Mat. USSR-Sb., 10 (1970), pp. 217–243.
[16] P. D. Lax and X.-D. Liu, *Solution of two-dimensional Riemann problems of gas dynamics by positive schemes*, SIAM J. Sci. Comput., 19 (1998), pp. 319–340.
[17] W. B. Lindquist, *The scalar Riemann problem in two spatial dimensions: Piecewise smoothness of solutions and its breakdown*, SIAM J. Math. Anal., 17 (1986), pp. 1178–1197.
[18] W. B. Lindquist, *Construction of solutions for two-dimensional Riemann problems*, Comput. Math. Appl. Part A, 12 (1986), pp. 615–630.
[19] C. W. Schulz-Rinne, *Classification of the Riemann problem for two-dimensional gas dynamics*, SIAM J. Math. Anal., 24 (1993), pp. 76–88.
[20] C. W. Schulz-Rinne, J. P. Collins, and H. M. Glaz, *Numerical solution of the Riemann problem for two-dimensional gas dynamics*, SIAM J. Sci. Comput., 14 (1993), pp. 1394–1414.
[21] W. Sheng and T. Zhang, *The Riemann Problem for the Transportation Equations in Gas Dynamics*, Mem. Amer. Math. Soc. 137, AMS, Providence, RI, 1999.
[22] D. Tan and T. Zhang, *Two-dimensional Riemann problem for a hyperbolic system of nonlinear conservation laws* (I): *Four-J cases*, J. Differential Equations, 111 (1994), pp. 203–254.
[23] D. Tan and T. Zhang, *Two-dimensional Riemann problem for a hyperbolic system of nonlinear conservation laws* (II): *Initial data consists of some rarefaction*, J. Differential Equations, 111 (1994), pp. 255–283.
[24] D. Tan, T. Zhang, and Y. Zheng, *Delta-shock waves as limits of vanishing viscosity for hyperbolic system of conservation laws*, J. Differential Equations, 112 (1994), pp. 1–32.
[25] B. Temple, *Global solution of the Cauchy problem for a class of* 2×2 *nonstrictly hyperbolic conservation laws*, Adv. in Appl. Math., 3 (1982), pp. 335–375.
[26] Y. Val'ka, *Discontinuous solutions of a multidimensional quasilinear equation (numerical experiments)*, U.S.S.R. Comput. Math. and Math. Phys., 8 (1968), pp. 257–264.
[27] A. I. Vol'pert, *The spaces BV and quasilinear equations*, J. Mat. USSR-Sb., 2 (1967), pp. 225–267.
[28] D. H. Wagner, *The Riemann problem in two space dimensions for a single conservation law*, SIAM J. Math. Anal., 14 (1983), pp. 534–559.
[29] S. Yang and T. Zhang, *The MmB difference solutions to the Riemann problem for a* 2*-D hyperbolic system of nonlinear conservation laws*, Impact Comput. Sci. Engrg., 3 (1991), pp. 146–180.
[30] P. Zhang, J. Li, and T. Zhang, *On two-dimensional Riemann problem for pressure-gradient equations of the Euler system*, Discrete Contin. Dynam. Systems, 4 (1998), pp. 609–634.
[31] P. Zhang and T. Zhang, *Generalized characteristic analysis and Guckenheimer structure*, J. Differential Equations, 152 (1999), pp. 409–430.
[32] T. Zhang and G. Chen, *Some fundamental concepts about systems of two spatial dimensions conservation laws*, Acta Math. Sinica, 6 (1986), pp. 463–474.
[33] T. Zhang and Y. Zheng, *Two dimensional Riemann problem for a single conservation law*, Trans. Amer. Math. Soc., 312 (1989), pp. 589–619.
[34] T. Zhang and Y. Zheng, *Conjecture on structure of solutions of the Riemann problem for two-dimensional gas dynamics systems*, SIAM J. Math. Anal., 21 (1990), pp. 593–630.
[35] T. Zhang and Y. Zheng, *Exact spiral solutions of the two dimensional compressible Euler equations*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 117–133.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 359–384

THE 2-DIMENSIONAL RIEMANN PROBLEM FOR A 2×2 HYPERBOLIC CONSERVATION LAW II. ANISOTROPIC MEDIA*

WOONJAE HWANG† AND W. BRENT LINDQUIST‡

Abstract. We construct the solutions to a two-dimensional (2-D) Riemann problem for a 2×2 hyperbolic nonlinear system which models polymer flooding in an anisotropic oil reservoir. The construction demonstrates the importance of the shock, rarefaction, and contact "base points" and "base curves" in the determination of the solutions for 2-D Riemann problems. In particular, we establish some new relations between these. While specific details of the base points and curves are applicable only to this model, the existence of the curves and the existence of relationships between these curves are general features to be exploited for any hyperbolic system.

Key words. Riemann problems, hyperbolic systems, conservation laws

AMS subject classifications. Primary, 35C05; Secondary, 35L65

PII. S0036141001396643

1. Introduction. Early work [1, 17, 8, 9, 21, 22] in two dimensions developed fundamental concepts for Riemann problem solutions for scalar equations. A by-product of this work was the development of a constructive technique for two-dimensional (2-D) Riemann problem solutions. Elements of this technique are still being refined concurrent with its use to obtain and examine solutions for systems where, in contrast to scalar hyperbolic equations, no general existence and uniqueness theory exists in multiple space dimensions. Spearheaded largely by the work of T. Zhang and collaborators, a body of work [7, 10, 11, 12, 13, 14, 18, 19, 23, 24] has since developed for the solution of 2-D Riemann problems for systems, specifically for the Euler system modeling gas dynamics and for simpler gas-dynamics-like models. In this work we consider a 2×2 system which is a 2-D generalization of the one-dimensional (1-D) Keyfitz–Kranzer–Isaacson–Temple model [4, 5, 15], which has been used in studies of material elasticity and flow in porous media. In addition to the particular solutions developed for this system, our emphasis here is on the constructive technique, in particular the fundamental role played by rarefaction, shock, and contact "base points" and "base curves," and the relationship between these elements.

In this and part I of this paper [3] we consider the 2×2 system

$$s_t + f^A(s,c)_x + f^B(s,c)_y = 0, \tag{1.1}$$

$$(cs)_t + (cf^A(s,c))_x + (cf^B(s,c))_y = 0. \tag{1.2}$$

An attractive feature of this model is that, regardless of the form of the flux functions f^A and f^B, the wave family associated with the variable c is always linear, i.e., produces only states of constant c-value separated by contact discontinuities. The other family, associated with the variable s, is identical to the scalar family obtained from (1.1) with c held constant.

*Received by the editors October 18, 2001; accepted for publication (in revised form) April 25, 2002; published electronically October 31, 2002. This work was supported by the Applied Mathematics Subprogram of the U.S. Department of Energy, grant DE-FG02-90ER25084.
http://www.siam.org/journals/sima/34-2/39664.html
†Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609-2280 (woonjae@wpi.edu).
‡Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794-3600 (lindquis@ams.sunysb.edu).

The presence of the linear c-family is important in guaranteeing solution uniqueness within a regularity class of functions. Recall [8] that initial data for a 2-D Riemann problem consists of piecewise constant data given on a finite number of wedges centered at the origin. As noted in part I of this paper [3], the system (1.1), (1.2) has no mechanism to remove (create) any constant concentration value $\bar{c}$ which is (is not) in the initial data. Consequently, only regions of constant concentration value exist in the self-similar Riemann problem solution, and these must correspond in an identifiable fashion with each region of constant concentration value in the initial data. These constant concentration regions must be delineated by contact discontinuities of the c-family. Furthermore, the solution within each region of constant concentration is determined by the scalar equation (1.1) for which uniqueness and existence of solutions within the function classes of bounded variation [16] and piecewise smooth [6] have been determined. Thus the only nonuniqueness possible in the solutions may arise from the pattern of contact discontinuities in a solution. As we show in section 4, the Rankine–Hugoniot conditions for the degenerate family lead to a unique pattern of contact discontinuity. The construction method used for the 2-D solutions employs the existence and uniqueness conditions of [6]. Thus within the class of piecewise smooth solutions (in the sense of Kruzkov), the solutions we generate are unique.

Complexity of the solution is governed by the form of the flux functions f^A and f^B and by the number of initial data wedges considered. To minimize complexity, we choose particular forms for the flux functions and limit the number of wedges. Our choice of flux function form is determined by application to flow in porous media. We consider the functions

$$\begin{aligned} f^A(s,c) &= s^2[1 + A(1-c)(1-s)], \qquad 0 < A < 1/2, \\ f^B(s,c) &= s^2[1 + B(1-c)(1-s)], \qquad 0 < B < 1/2, \end{aligned} \tag{1.3}$$

where the physical state variables are

$$\begin{aligned} s &= \text{water saturation}, & 0 \le s \le 1, \\ c &= \text{concentration of polymer}, & 0 \le c \le 1. \end{aligned}$$

This system models polymer flooding of an oil reservoir [4]. In a polymer flood, a small amount of polymer is added to the water to increase the sweep efficiency of oil production. The model assumes the polymer is completely miscible in the water phase and undergoes no mass transfer into the oil phase. In part I of this paper [3] we have developed the solution of this model for the case $A \equiv B$ (i.e., $f^A \equiv f^B$), which is applicable to isotropic media.

In this paper, we develop solutions for the case $A \neq B$. The directional fluxes f^A and f^B differ in those terms of order higher than s^2, as may be appropriate for anisotropic media. In order to restrict the complexity introduced by compound waves in the s-family we simplify to a convex form for the flux functions. The restriction $0 < A < 1/2$ guarantees that the x-direction flux function f^A remains convex for allowed values of $s \in [0,1]$ and $c \in [0,1]$, similarly for the y-direction flux f^B. The following lemma impacts the form of the effective flux function $f_s^{\hat{n}} \equiv \hat{n} \cdot (f_s^A, f_s^B)$ governing flow in an arbitrary 2-D direction $\hat{n}$.

LEMMA 1.1 (see [22]). *For fixed c, if $f_{ss}^A \neq 0, f_{ss}^B \neq 0$, and $\frac{\partial}{\partial s}(f_{ss}^A / f_{ss}^B) \neq 0$ for any s then, for any $\hat{n}$, $f^{\hat{n}}$ has at most one inflection point.*

Lemma 1.1 holds under either the restriction $0 < B < A < 1/2$ or $0 < A < B < 1/2$; without loss of generality we will consider the case $0 < A < B < 1/2$.

The 2-D Riemann problem studies for gas dynamics models have involved initial data on four wedges corresponding to the four quadrants in the plane, as this is pertinent to typical data encountered in 2-D finite difference methods. For general

Riemann data, n waves emerge from an initial discontinuity for an $n \times n$ system. The work on gas dynamics models has typically added an additional restrictive assumption on the initial data values to ensure that only a single wave evolves from each initial discontinuity. In this paper, we relax the single wave restriction; however, we do restrict the initial data to two wedges defined by the angles $\pi/2$ and π (Figure 4.1(a)). As the smaller wedge agrees with one of the quadrants in the plane, we refer to this as a single quadrant Riemann problem.

To analyze system (1.1), (1.2) it is convenient to introduce the change of variables $s, c \to s, b \equiv sc$. The eigenvalues, right eigenvectors, and Riemann invariants for the two families of waves are, respectively,

$$\begin{array}{rlll} \lambda^s = f_s^{\hat{n}}, & \text{where} & f_s^{\hat{n}} \equiv \hat{n} \cdot (f_s^A, f_s^B), & \\ \lambda^c = g^{\hat{n}}, & \text{where} & g^{\hat{n}} \equiv \hat{n} \cdot (g^A, g^B), & g^\alpha \equiv f^\alpha / s, \quad \alpha = A, B, \\ r^s = (s, b), & & r^c = (-g_b^{\hat{n}}, g_s^{\hat{n}}), & \\ W^s = c, & & W^c = g^{\hat{n}}. & \end{array}$$

Here, $\hat{n} \equiv (\mu, \nu)$ denotes a direction vector and subscripts represent partial differentiation.

The solution construction method developed in [1, 9, 17, 22] relies heavily on the existence of so-called base points and curves. We review these in section 2 and establish some new relations between these. While specific details of the base points and curves are applicable only to this model, the existence of the curves and the existence of relationships between these curves are general features to be exploited for any hyperbolic system. In particular, the geometrical nature of the Kruzkov entropy condition (Figure 2.3) and the relationships between the shock and rarefaction base curves place regional restrictions on the allowed s-family waves. This is discussed for this model in section 3. A topological classification of the solutions and a discussion of the generic solution form for the single quadrant Riemann problem for our system is presented in section 4. In section 5 we present the unique canonical Riemann problem solution in each class.

2. Base curves. As the solution to a Riemann problem is self-similar, direct construction of a 2-D Riemann problem solution is done in the plane defined by the self-similar coordinates $\xi = x/t$, $\eta = y/t$ (i.e., in the plane $t = 1, x, y$). Positioning of shock, rarefaction, and contact discontinuity waves in this plane is related to the existence of rarefaction and shock base curves and contact base points. Furthermore, shock and rarefaction waves are each classifiable into two types [22], and possible interactions among the types are limited.

The existence of base points and curves, the existence of relations between them, and the shock and rarefaction wave classifications are general. However, the form of base curves and position of base points depend on the particular choice of flux functions. The flux functions in the model studied here have the following characteristics. For any fixed value of $c \in [0, 1]$ (recall we have chosen $A < B$),

(2.1) $f^\alpha(0, c) = 0, \quad f^\alpha(1, c) = 1, \quad \alpha = A, B;$

(2.2) $f^\alpha(s, c)$ is convex (i.e., $f_s > 0, \quad f_{ss} > 0$ for $s \in [0, 1]$), $\quad \alpha = A, B;$

(2.3) $f_s^A < f_s^B$ for $s \in [0, s_{AB}), \quad f_s^B < f_s^A$ for $s \in (s_{AB}, 1];$

(2.4) $g^\alpha(0, c) = 0, \quad g^\alpha(1, c) = 1, \quad \alpha = A, B;$

(2.5) $g^A(s, c) < g^B(s, c)$ for $s \in [0, 1];$

(2.6) $g^\alpha(s, c) \le f_s^\alpha(s, c)$ for $s \in [0, 1], \quad \alpha = A, B,$ with equality holding only for $s = 0$.

For any fixed value of $s \in [0,1]$,

$$f^\alpha(s,c_1) > f^\alpha(s,c_2) \text{ for } c_1 < c_2, \quad \alpha = A,B; \tag{2.7}$$

$$g^\alpha(s,c_1) > g^\alpha(s,c_2) \text{ for } c_1 < c_2, \quad \alpha = A,B. \tag{2.8}$$

2.1. Rarefaction base curves and wave classification. Consider the scalar equation (1.1) with $c = c_0$ (c_0 a constant).

$$s_t + f^A(s,c_0)_x + f^B(s,c_0)_y = 0. \tag{2.9}$$

Under the change of variables $\xi = x/t$, $\eta = y/t$, (2.9) has the self-similar form

$$-\xi s_\xi - \eta s_\eta + f^A(s,c_0)_\xi + f^B(s,c_0)_\eta = 0, \tag{2.10}$$

where now $s = s(\xi,\eta)$. For $s \in C^1$, (2.10) becomes

$$\left(f^A_s(s,c_0) - \xi\right)s_\xi + \left(f^B_s(s,c_0) - \eta\right)s_\eta = 0 \tag{2.11}$$

whose characteristic form is given by

$$\begin{aligned} d\eta(\xi)/d\xi &= (f^B_s(s,c_0) - \eta)/(f^A_s(s,c_0) - \xi), \\ ds(\xi,\eta(\xi))/d\xi &= 0. \end{aligned} \tag{2.12}$$

From (2.12), the characteristic lines are defined by

$$\frac{\eta - f^B_s(s,c_0)}{\xi - f^A_s(s,c_0)} = \text{const}, \qquad s = \text{const}. \tag{2.13}$$

Remark 2.1. From (2.13) we note that an s-family level curve with $s = s_1$, $c = c_0$ is a straight line segment whose tangent passes through the *characteristic base point* $f_s(s_1;c_0)$ in the ξ,η plane having coordinates

$$f_s(s_1;c_0) \equiv (f^A_s(s_1,c_0), f^B_s(s_1,c_0)), \tag{2.14}$$

as illustrated in Figure 2.1(a). We refer to the curve

$$f_s(s;c_0) \equiv (f^A_s(s,c_0), f^B_s(s,c_0)), \qquad 0 \le s \le 1, \tag{2.15}$$

as the *rarefaction base curve for* c_0.

As a consequence of Remark 2.1, (i) if a single point on a characteristic in a rarefaction wave is known the characteristic can be extended locally as a straight line segment, or (ii) if the direction of a characteristic wave is known the wave can be locally located (as a straight line segment) in the (ξ,η) plane.

For a model with the characteristics (2.1)–(2.2) it is easy to show that the rarefaction base curve is a monotonic increasing, concave curve segment in the (ξ,η) plane extending between the points $(\xi = 0, \eta = 0)$ and $(\xi = f^A_s(1,c), \eta = f^B_s(1,c))$. For the fluxes (1.3) used here, the value s_{AB} in (2.3) is independent of the value of c, A, or B; thus all rarefaction base curves pass through the point $(f^A_s(s_{AB},c), f^B_s(s_{AB},c)) = (4/3, 4/3)$. Two example rarefaction base curves are shown in Figure 2.1(b). The upper termination point $(2 - A(1-c), 2 - B(1-c))$ of any rarefaction base curve must lie within the triangular region delimited by the dotted lines.

Rarefaction waves can be classified [20] according to the direction of the gradient of s across the wave relative to the direction toward the base curve.

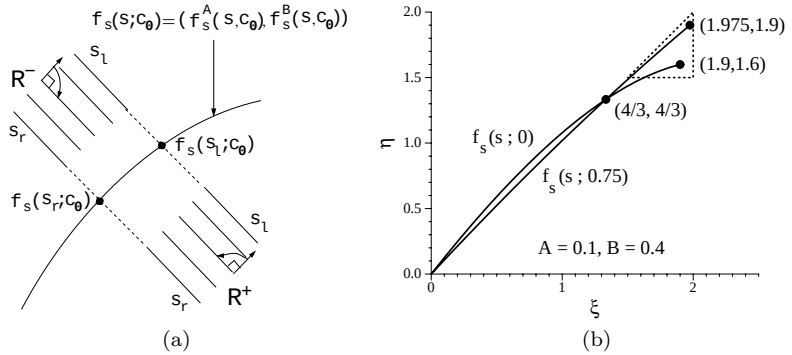


FIG. 2.1. (a) *A rarefaction base curve* $f_s(s) \equiv (f_s^A(s), f_s^B(s))$ *and rarefaction waves* R^+, R^- $(s_l > s_r)$. (b) *Example plots of two rarefaction base curves.*

DEFINITION 2.1 (see [20]). *Classify a rarefaction wave as* R^+ (R^-) *if* $\nabla_{\xi,\eta} s$ *and the direction toward the base curve of the characteristic lines of the wave form a right- (left-) hand system.*

R^+ and R^- rarefaction waves are indicated in Figure 2.1(a). If (s_l, c) and (s_r, c) are the bounding characteristics of a rarefaction wave, the wave will be labeled $R^+_{l,r}$ or $R^-_{l,r}$ as appropriate.

2.2. Shock base curves and wave classification. From (2.10) we have the Rankine–Hugoniot condition for a piecewise smooth shock curve $\eta = \eta(\xi)$,

$$\frac{d\eta}{d\xi} = \frac{\eta - \sigma^B(s_l, s_r; c_0)}{\xi - \sigma^A(s_l, s_r; c_0)}, \tag{2.16}$$

where

$$\sigma^\alpha(s_l, s_r; c_0) \equiv \frac{f^\alpha(s_l, c_0) - f^\alpha(s_r, c_0)}{s_l - s_r}, \qquad \alpha = A, B. \tag{2.17}$$

Remark 2.2. From (2.16) we see that any s-family shock point D in the (ξ, η) plane separating (s_l, c_0) and (s_r, c_0) lies on a curve segment whose tangent at D passes through the *shock base point* $\sigma(s_l, s_r; c_0)$ in the ξ, η plane having coordinates

$$\sigma(s_l, s_r; c_0) \equiv (\sigma^A(s_l, s_r; c_0),\ \sigma^B(s_l, s_r; c_0)). \tag{2.18}$$

The notations $\sigma(s_l, s_r; c)$ and $\sigma(s_r, s_l; c)$ specify the same base point. The curve of base points

$$\sigma(s, s_r; c_0) \equiv (\sigma^A(s, s_r; c_0),\ \sigma^B(s, s_r; c_0)), \qquad 0 \le s \le 1, \tag{2.19}$$

is denoted the *shock base curve* for the state (s_r, c_0).

Shock base points are used analogously to rarefaction base points to locally position shocks involving the states s_r, c_0 and s_l, c_0. For our model with $A < B$ any shock base curve $\sigma(s, s_r; c_0)$ is monotonic increasing and concave in the (ξ, η) plane. A shock base curve $\sigma(s, s_r; c_0)$ is sketched in Figure 2.2(a).

DEFINITION 2.2. *For a point* D *on a shock separating* s_l *and* s_r*, the direction of the normal vector* $\hat{n} \equiv (\mu, \nu)$ *at* D *is defined as pointing from the side having larger to the side having smaller value of* s*; i.e., if* $s_l > s_r$*,* $\hat{n}$ *points from* s_l *to* s_r*.*

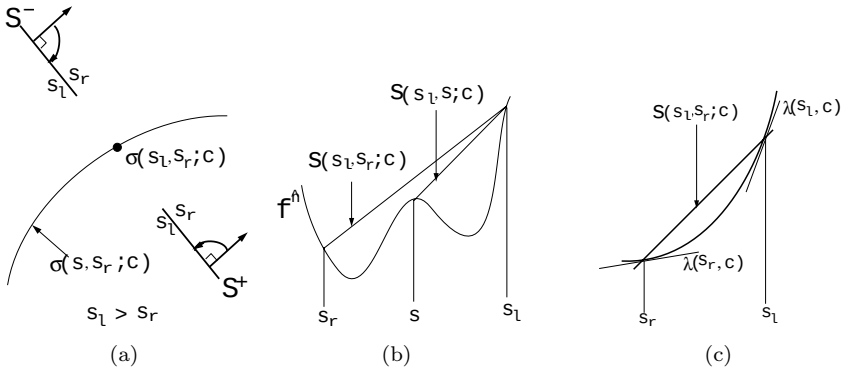


FIG. 2.2. (a) *Illustration of the shock base curve* $\sigma(s, s_r; c)$ *and* S^+ *and* S^- *shock waves. Sketch of the Kruzkov entropy condition for* (b) *general and* (c) *convex flux functions.* $S(\cdot)$ *and* $\lambda(\cdot)$ *refer to the slopes of indicated lines.*

Let

$$\begin{aligned} S(s_l, s_r; c) &\equiv \hat{n} \cdot (\sigma^A(s_l, s_r; c), \sigma^B(s_l, s_r; c)) \\ &= \mu\sigma^A(s_l, s_r; c) + \nu\sigma^B(s_l, s_r; c). \end{aligned} \tag{2.20}$$

The Kruzkov entropy condition for a shock whose normal has the direction (μ, ν) is given by

$$S(s_l, s_r; c) \le S(s_l, s; c) \quad \text{for all} \quad s \in (s_r, s_l). \tag{2.21}$$

This entropy condition for a flux function $f^{\hat{n}}(s, c)$ of general shape is illustrated in Figure 2.2(b); it is the familiar Oleinik construction. If $f^{\hat{n}}$ is convex over the domain (s_r, s_l), the entropy condition reduces to

$$\lambda(s_r, c) < S(s_l, s_r; c) < \lambda(s_l, c), \tag{2.22}$$

i.e.,

$$\hat{n} \cdot (f_s^A, f_s^B)\Big|_{s_r} < \hat{n} \cdot (\sigma^A, \sigma^B) < \hat{n} \cdot (f_s^A, f_s^B)\Big|_{s_l}. \tag{2.23}$$

This convex form of the entropy condition is illustrated in Figure 2.2(c) and is the well-known Lax entropy condition. As shown in Figure 2.3, the Lax entropy inequality (2.23) is equivalent to comparing the lengths of the projections on the direction $\hat{n}$ of the vectors from the origin of the ξ, η plane to the respective rarefaction and shock base points $f_s(s_r; c), \sigma(s_l, s_r; c)$, and $f_s(s_l; c)$.

A shock wave can be classified [22] according to the relative direction of the shock normal and tangent vectors at each point.

DEFINITION 2.3 (see [22]). *A shock wave is classified as* S^+ *(*S^-*) if at each point the normal and tangent (pointing towards the shock base point) vectors of the shock form a right-hand (left-hand) system.*

The two shock types are also illustrated in Figure 2.2(a). If (s_l, c) and (s_r, c) are the states bounding the shock, the shock will be labeled $S_{l,r}^+$ or $S_{l,r}^-$ as appropriate.

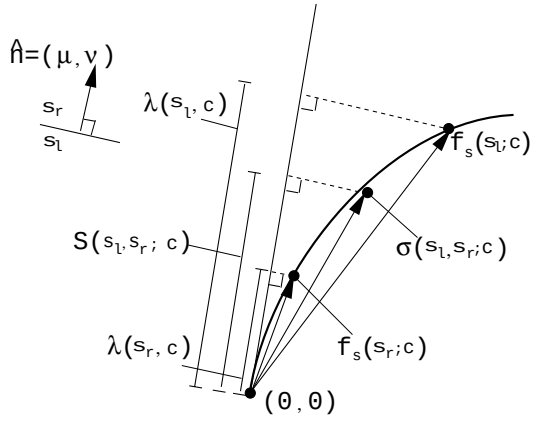


FIG. 2.3. *Geometrical realization in the ξ, η plane of the Kruzkov entropy condition for convex f.*

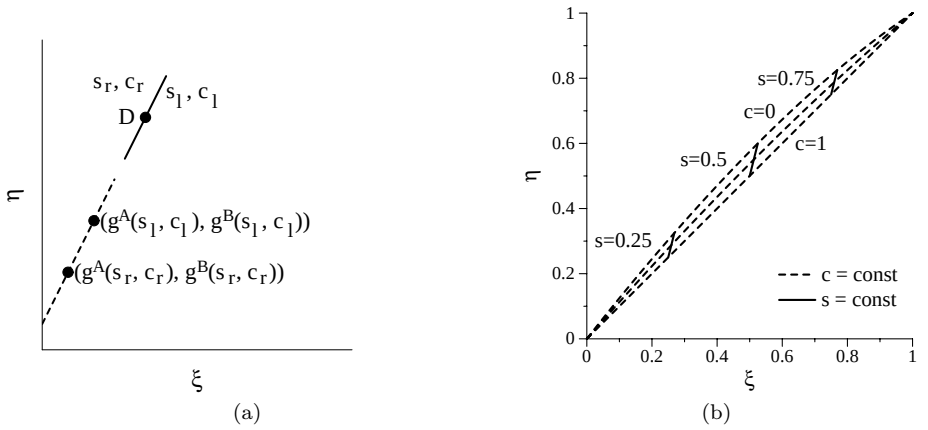


FIG. 2.4. (a) *The tangent (dotted line) to a contact discontinuity point D must pass through two appropriate contact base points $(g^A(s_l, c_l), g^B(s_l, c_l))$ and $(g^A(s_r, c_r), g^B(s_r, c_r))$ as illustrated.* (b) *Illustration of coordinate curves which cover the region of contact base points. Here $s = 0.5$ refers to the coordinate curve $(g^A(0.5, c), g^B(0.5, c))$, while $c = 0$ refers to the coordinate curve $(g^A(s, 0), g^B(s, 0))$, etc.*

2.3. Contact discontinuity base points. Let D be a point on a c-family contact discontinuity curve segment separating the states (s_l, c_l) and (s_r, c_r). The Rankine–Hugoniot condition across a smooth c-family contact discontinuity curve having normal vector $\hat{n} = (\mu, \nu)$ at D is

$$g^{\hat{n}}(s_l, c_l) = g^{\hat{n}}(s_r, c_r), \tag{2.24}$$

where $g^{\hat{n}} \equiv \mu g^A + \nu g^B$.

Remark 2.3. As a consequence of (2.24) a contact discontinuity point D lies on a curve segment whose tangent must pass through the two *contact base points*,

$$(g^A(s_l, c_l), g^B(s_l, c_l)) \quad \text{and} \quad (g^A(s_r, c_r), g^B(s_r, c_r)), \tag{2.25}$$

as illustrated in Figure 2.4(a).

For a model with the characteristics (2.4), (2.5) any contact base point must lie within a region $\Omega_{A,B}$ of the ξ, η plane bounded between the concave curve segments $(g^A(s,1), g^B(s,1))$ and $(g^A(s,0), g^B(s,0))$, both of which terminate at the points (0, 0) and (1, 1). An example region $\Omega_{0.1,0.4}$ for the model studied here is shown in Figure 2.4(b). In this case $(g^A(s,1), g^B(s,1))$ is the straight line segment between (0,0) and (1,1). The set of curve segments $(g^A(s, c = \text{const}),\ g^B(s, c = \text{const}))$ and the set $(g^A(s = \text{const}, c),\ g^B(s = \text{const}, c))$ form a coordinate system over this region. (Thus the contact base point $(g^A(s_l, c_l), g^B(s_l, c_l))$ occurs at the intersection of the $(g^A(s, c_l), g^B(s, c_l))$ and $(g^A(s_l, c), g^B(s_l, c))$ coordinate curves.) The $(g^A(s = \text{const}, c),\ g^B(s = \text{const}, c))$ coordinate curves are straight lines, with slope B/A. A few coordinate curves are also sketched in Figure 2.4(b).

2.4. Relation between rarefaction and shock base curves. For a fixed value of $c = c_0$ and any $s_1 \in (0,1)$, the shock and rarefaction base curves $\sigma(s_1, s; c_0)$ and $f_s(s; c_0)$ are related as indicated in the following two lemmas.

LEMMA 2.4.

$$\frac{d}{ds}\sigma(s_1, s; c_0) = \frac{1}{(s - s_1)}\big[f_s(s; c_0) - \sigma(s_1, s; c_0)\big]. \tag{2.26}$$

Proof. We prove (2.26) componentwise. Consider the ξ-component

$$\begin{aligned}\frac{d}{ds}\sigma^A(s_1, s; c_0) &\equiv \frac{d}{ds}\left(\frac{f^A(s, c_0) - f^A(s_1, c_0)}{s - s_1}\right)\\ &= \frac{1}{(s - s_1)}\left(f_s^A(s, c_0) - \frac{f^A(s, c_0) - f^A(s_1, c_0)}{s - s_1}\right)\\ &= \frac{1}{(s - s_1)}\big(f_s^A(s, c_0) - \sigma^A(s_1, s; c_0)\big).\end{aligned}$$

The computation is identical for the η-component. ∎

LEMMA 2.5.

$$\frac{d\sigma^B(s_1, s; c_0)}{d\sigma^A(s_1, s; c_0)} = \frac{f_s^B(s, c_0) - \sigma^B(s_1, s; c_0)}{f_s^A(s, c_0) - \sigma^A(s_1, s; c_0)}. \tag{2.27}$$

Proof. The proof follows immediately from Lemma 2.4 using the fact that s_1 and c_0 are constant, i.e.,

$$\frac{d\sigma^B(s_1, s; c_0)}{d\sigma^A(s_1, s; c_0)} = \frac{d\sigma^B(s_1, s; c_0)/ds}{d\sigma^A(s_1, s; c_0)/ds}. \qquad ∎$$

Remark 2.4. Two shock base curves $\sigma(s, s_r; c)$ and $\sigma(s_l, s; c)$ pass through each shock base point $\sigma(s_l, s_r; c)$. Lemma 2.5 states that the tangent to the shock base curve $\sigma(s, s_r; c)$ at the shock base point $\sigma(s_l, s_r; c)$ also passes through the rarefaction base curve at the point $f_s(s_l; c)$. Equivalently, the tangent to the shock base curve $\sigma(s_l, s; c)$ at the shock base point $\sigma(s_l, s_r; c)$ passes through the rarefaction base curve at the point $f_s(s_r; c)$. These tangent lines (dashed lines) are shown in Figure 2.5(a).

As a consequence of Remark 2.4, the rarefaction and shock base points $f_s(s_r; c)$ and $\sigma(s_r, s_r; c)$ are coincident and the shock $\sigma(s, s_r; c)$ and the rarefaction $f_s(s; c)$ base curves meet tangentially at this common point (i.e., at the parameter value $s = s_r$). This is also illustrated in Figure 2.5(a).

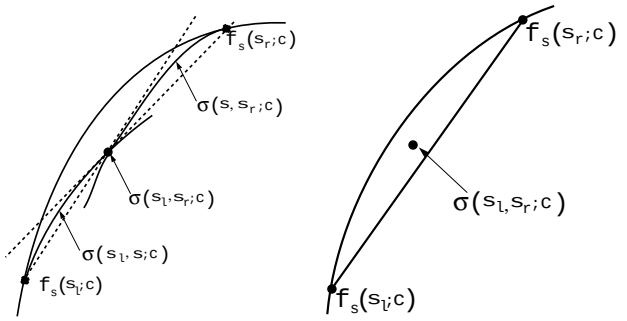


FIG. 2.5. *Illustration of* (a) *the relation between the rarefaction and shock base curves. The dashed lines indicate tangents to the two shock base curves at* $\sigma(s_r, s_l; c)$*;* (b) *the convex hull property of shock base points relative to the rarefaction base curve.*

Lemmas 2.4 and 2.5 and Remark 2.4 hold for any scalar equation with flux functions having sufficient continuity. The following remark is pertinent to the model studied here.

Remark 2.5. The rarefaction $f_s(s; c)$ and shock $\sigma(s_1, s; c)$ base curves have the same curvature sign. This property follows from Remark 2.4

LEMMA 2.6. *For our model, the shock base point* $\sigma(s_l, s_r; c)$ *lies within the convex hull of* $f_s(s; c)$ *between the rarefaction base points* $f_s(s_r; c)$ *and* $f_s(s_l; c)$ *as shown in Figure* 2.5(b). *We refer to this as the* convex hull *property of the rarefaction base curve relative to shock base points.*

Proof. As a consequence of Remark 2.5, the larger of the two angles between the two tangent lines in Remark 2.4 must be less than π. □

2.5. Relation between rarefaction and contact base coordinate curves. Following a proof analogous to that for Lemma 2.5, we have the fundamental relation between the rarefaction base curve for c_0 and the $c = c_0$ contact base coordinate curve.

LEMMA 2.7.

$$\frac{dg^B(s, c_0)}{dg^A(s, c_0)} = \frac{f_s^B(s, c_0) - g^B(s, c_0)}{f_s^A(s, c_0) - g^A(s, c_0)}. \tag{2.28}$$

Note that this lemma depends only on the definition $g(s, c) \equiv f(s, c)/s$ and not on the form of $f(\cdot)$.

Remark 2.6. Lemma 2.7 states that the tangent to the contact base point coordinate curve $(g^A(s, c_0), g^B(s, c_0))$ at the contact base point $(g^A(s_1, c_0), g^B(s_1, c_0))$ passes through the rarefaction base curve at the base point $f_s(s_1; c_0)$.

Remark 2.7. Further, in this model it is easy to show the following.

1. Both the rarefaction base curve $f_s(s; c_0)$ and contact base point coordinate curve $(g^A(s, c_0), g^B(s, c_0))$ have the same curvature sign. For $A < B$ the curvature is negative.
2. For $A < B$ and $s \in (0, 1]$, the coordinate curve $(g^A(s, c_0), g^B(s, c_0))$ lies below the rarefaction base curve $f_s(s; c_0)$.
3. The base points $(g^A(0, c_0), g^B(0, c_0))$ and $f_s(0; c_0)$ are identical, and the contact base coordinate curve and the rarefaction base curve meet tangentially at this common point (which is in fact the point $(\xi = 0, \eta = 0)$).

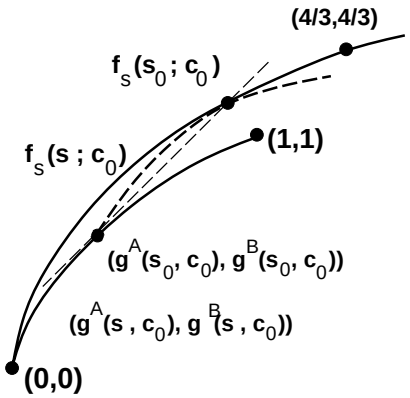


FIG. 2.6. *Illustration of the relationship between the rarefaction base curve $f_s(s;c_0)$ (upper), the contact base coordinate curve $(g^A(s,c_0), g^B(s,c_0))$ (lower), and the shock base curve $\sigma(s_1,s;c_0)$ (dashed). The straight dashed line is tangent to the contact base coordinate curve at $(g^A(s,c_0), g^B(s,c_0))$.*

Figure 2.6 shows a rarefaction base curve and the corresponding contact base point coordinate curve.

2.6. Relation between shock base curves and contact base points. The following lemma establishes a useful relationship between shock base curves and contact base points.

LEMMA 2.8.

$$\left.\begin{array}{ll} g^\alpha(s_0,c_0) = \sigma^\alpha(s_0,0;c_0), & \\ g^\alpha(s_0,c_0) < \sigma^\alpha(s_0,s;c_0), & s\in(0,1], \end{array}\right\} \qquad \alpha = A,B. \tag{2.29}$$

Proof. The equality in the lemma follows from the definitions of g^α and σ^α. The inequality is easily shown geometrically given that the function $f^\alpha(s,c_0)$ is convex. ∎

From Lemma 2.8 we see that the contact base point $(g^A(s_0,c_0), g^B(s_0,c_0))$ is the "lower" termination point of the shock base curve $\sigma(s_0,s;c_0)$. This relationship is also sketched in Figure 2.6.

3. Entropy restrictions for *s*-family shocks. The geometrical nature of the entropy condition (Figure 2.3) and the relationships between the shock and rarefaction base curves place the following regional restrictions on the allowed s-family waves in the ξ,η plane.

Consider a point D on a curve of discontinuity separating states s_l, c and s_r, c with $s_l > s_r$. Assume at D that the normal and tangent (pointing toward the appropriate shock base point) form a right-handed system. If D lies in the lower right, wedge shaped region of the ξ,η plane labeled $S^+_{l,r}$ in Figure 3.1(a) (this region is centered on the shock base point $\sigma(s_l,s_r;c)$, is bounded above by the half line starting from this shock base point and passing through the rarefaction base point $f_s(s_l;c)$, and is bounded to the left by the half line starting from the shock base point and passing through the rarefaction base point $f_s(s_r;c)$), then the local solution to this Riemann problem is an s-family shock of type $S^+_{l,r}$.

It is an easy geometrical exercise to verify the entropy condition for this shock. The verification is sketched in Figure 3.2(a). The projection lengths (2.22) are seen to have the correct ordering along the normal direction $\hat{n}$.

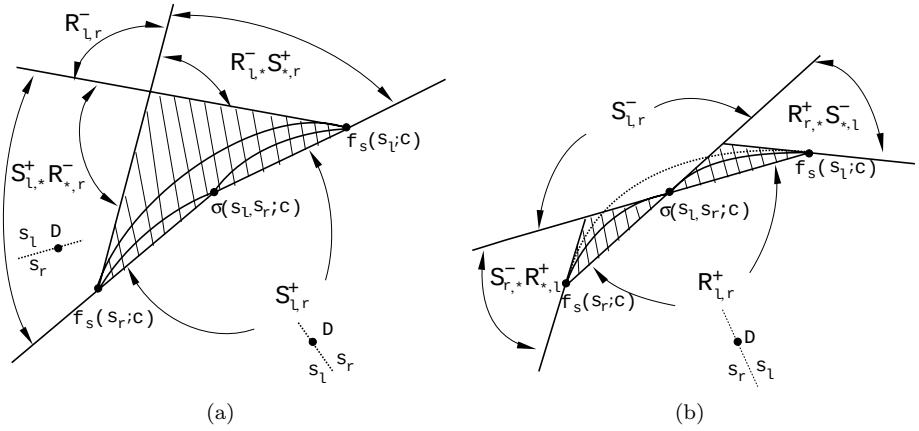


FIG. 3.1. *Regional entropy restrictions for s-family waves. Here* $s_l > s_r$.

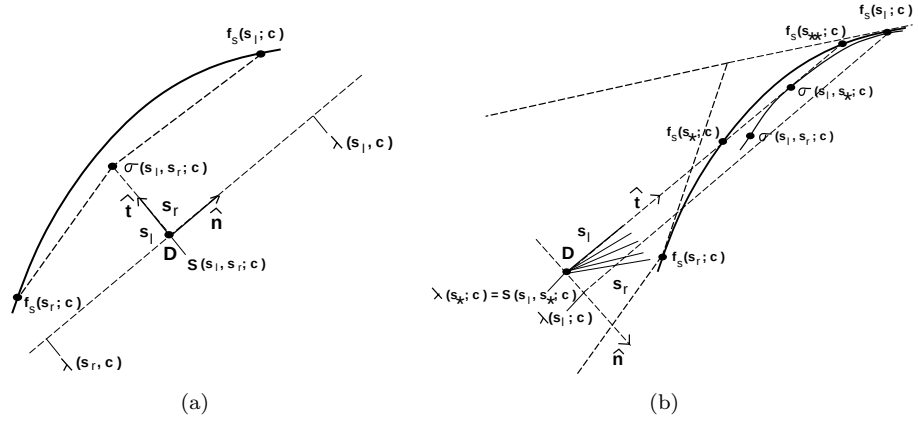


FIG. 3.2. (a) *Detail verifying the Lax entropy condition for an* S^+ *shock in the* $S^+_{l,r}$ *region.* (b) *Detail verifying the Lax entropy condition for the composite* S^+R^- *wave shock in the* $S^+_{l,*}R^-_{*,r}$ *region.*

If D lies in the region labeled $S^+_{l,*}R^-_{*,r}$ (lower left) in Figure 3.1(a), with the orientation of s_l and s_r ($s_l > s_r$) as indicated, the local Riemann problem solution will consist of a composite wave consisting of an $S^+_{l,*}$ shock and an $R^-_{*,r}$ rarefaction fan. Here $*$ denotes the intermediate state common to one side of the shock and the rarefaction fan. The region is bounded by the three dark dashed lines shown in Figure 3.2(b). (The lower dashed line is the extension of the line segment from the shock base point $\sigma(s_l, s_r; c)$ to the rarefaction base point $f_s(s_r; c)$. The middle dashed line is tangent to the rarefaction base curve at $f_s(s_r; c)$. The upper is tangent to the rarefaction base curve at $f_s(s_l; c)$.)

LEMMA 3.1. *There is a unique base point* $\sigma(s_l, s_*; c)$, *with* $s_* < s_l$, *on the shock base curve* $\sigma(s_l, s; c)$ *such that the line segment from* D *to* $\sigma(s_l, s_*; c)$ *is tangent to the shock base curve. This line also passes through a rarefaction base point* $f_s(s_*; c)$ *such that* $\sigma(s_l, s_*; c)$ *lies in the convex hull of* $f_s(s_*; c)$ *and* $f_s(s_l; c)$.

Proof. The existence of the unique shock base point $\sigma(s_l, s_*; c)$ follows by requiring D to remain within the region bounded by the indicated lines and by noting that the shock base curve is concave down. The relations between rarefaction and shock base curves noted in section 2.4 imply that this tangent line passes through the rarefaction base curve $f_s(s; c)$ twice, at base points $f_s(s^*; c)$ and $f_s(s^{**}; c)$ as indicated in Figure 3.2(b). However, only for the smaller (s^*) of the values s^* and s^{**} does $\sigma(s_l, s_*; c)$ lie in the convex hull of $f_s(s; c)$ between $f_s(s_l; c)$ and $f_s(s_*; c)$. ∎

Thus the unique local entropy solution of the Riemann problem at D consists of a shock between the state s_l and the state s_* followed by a rarefaction wave $R^-_{*,r}$. The tangent (and hence normal) to the shock direction at D is determined by the shock base point $\sigma(s_l, s_*; c)$. The characteristics in the rarefaction fan must point to the appropriate rarefaction base points corresponding to values of s lying in the range $[s_*, s_r]$ as sketched in the Figure 3.2(b). Again it is easy to geometrically verify the entropy condition

$$\hat{n} \cdot (f_s^A, f_s^B)\Big|_{s_*} = \hat{n} \cdot (\sigma^A, \sigma^B)\Big|_{l,*} < \hat{n} \cdot (f_s^A, f_s^B)\Big|_{s_l} \tag{3.1}$$

for the $S^+_{l,*}$ shock, as illustrated in Figure 3.2(b).

If D lies in the region labeled $R^-_{l,r}$ (upper left) in Figure 3.1(a), the local Riemann problem solution will consist of an $R^-_{l,r}$ rarefaction fan. If D lies in the region labeled $R^-_{l,*}S^+_{*,r}$ (upper right) in Figure 3.1(a), the local Riemann problem solution will consist of a composite wave consisting of an $R^-_{l,*}$ rarefaction fan followed by an $S^+_{*,r}$ shock. The detailed analysis of the composite wave in this region follows analogously to that for the $S^+_{l,*}R^-_{*,r}$ region.

If s_l and s_r are switched so that now the normal and tangent vectors at D form a left-handed system, four different solution regions in the ξ, η plane develop. These are indicated in Figure 3.1(b).

Note in Figure 3.1 that the occurrence of D in the shaded regions is not investigated. This shaded region is dynamic, depending on the left and right states on each side of the discontinuity. Solutions are constructed by "tracing waves in from infinity." We observe one of two events occurring as a shock point D is "traced in" and approaches the shaded region; either D separates two constant states, in which case the shock remains a straight line with unchanging entropy condition, or the change in saturation s across the shock weakens, and the dynamic shaded region shrinks in a manner that D never reaches it.

4. Solution classification and generic form. A 2-D Riemann problem solution can be partly classified by its solution at infinity in the (ξ, η) plane where it consists of noninteracting 1-D solutions. For the single quadrant problem Figure 4.1(a) investigated here, the solution form at infinity is sketched in Figure 4.1(b). In each direction the solution consists of a contact discontinuity followed by an s-family wave (which cannot be composite as the flux functions $f^A(s, c)$ and $f^B(s, c)$ are convex). It is therefore natural to label the solution behavior at infinity as $\mathrm{CW}_y\mathrm{CW}_x$, where W_y and W_x denote the s-family wave type (either S or R) in the y and x direction, respectively. Given s_1, c_1 there are four cases for the solution at infinity (labeled CSCS, CSCR, CRCS, and CRCR) determined by the location of s_2, c_2 in one of four regions in phase space. These four regions are indicated in Figure 4.2(a). The regions are bounded by the two curves $c^\alpha(s)$ defined implicitly by $g^\alpha(s, c) = g_1^\alpha$, where $g_1^\alpha \equiv f^\alpha(s_1, c_1)/s_1$, $\alpha = A, B$.

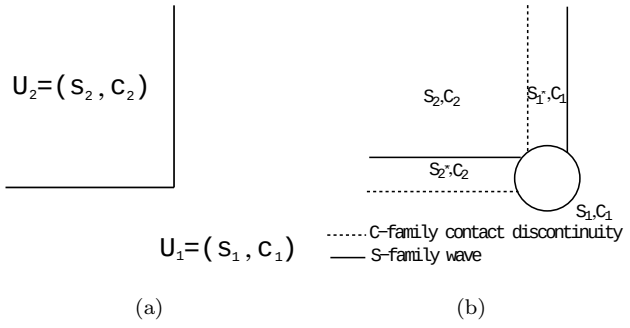


FIG. 4.1. (a) *The initial data for the single quadrant Riemann problem.* (b) *The solution at infinity.*

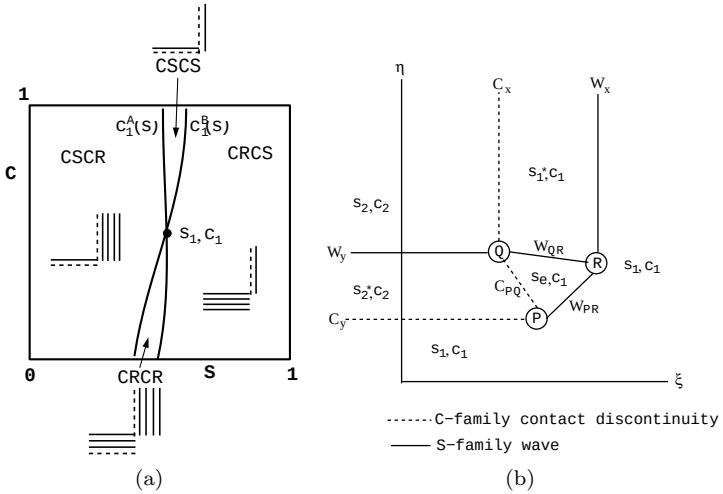


FIG. 4.2. (a) *Phase space classification of the solution at infinity.* (b) *The generic form of the single quadrant Riemann problem solution.*

As discussed earlier, the single quadrant Riemann problem solution must contain a "single" contact discontinuity wave which divides the solution into two separate scalar regions. This contact discontinuity provides the only mechanism for the $c_2 \to c_1$ transition required in the solution. Given this and the solution structure at infinity, all solutions must have the generic form shown in Figure 4.2(b). The existence of the constant state s_e, c_1 is a consequence of the required transition across the contact discontinuity C_{PQ}. Whereas Q and R may denote either an interaction point or interaction region in the solution, P is always an interaction point (since the contact discontinuities C_{PQ} and C_y must meet at P). In fact, P is the contact base point $(g^A(s_{2^*}, c_2), g^B(s_{2^*}, c_2))$.

In the rest of this section we discuss the interactions in regions Q and P. In section 5 we discuss the interaction in region R in the context of completing the entire solution.

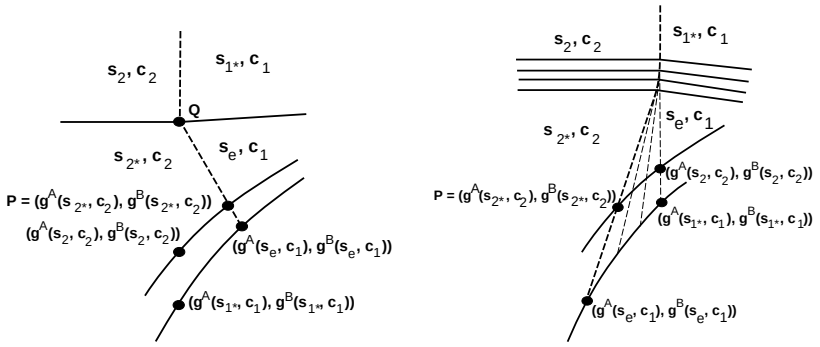


FIG. 4.3. *Illustration of construction in region Q for* (a) *CSCW$_x$ and* (b) *CRCW$_x$ cases. The lighter dashed lines in* (b) *illustrate the behavior of the tangents to the contact discontinuity.*

Interaction at *Q*. The wave W_{QR} is of the same type as W_y. The direction of W_{QR} is determined by the appropriate shock or rarefaction base curve. Our discussion therefore concentrates on the contact discontinuity segment C_{PQ} (Figure 4.2(b)) separating the constant states s_{2^*}, c_2 and s_e, c_1. Let $\hat{n} = (\mu, \nu)$ be the normal to C_{PQ} pointing from the side s_{2^*}, c_2 to s_e, c_1. Let $\theta_{\mu\nu} \equiv \tan^{-1}(\nu/\mu)$ be the angle associated with the normal.

LEMMA 4.1. *For solutions in which W_y is an s-family shock, $0 < \theta_{\mu\nu} < \pi/2$.*

Proof. When W_y is an s-family shock, Q in Figure 4.2(b) is a single point. Figure 4.3(a) illustrates the detail of the construction of this interaction. As W_y and C_y are horizontal lines, the y coordinate of point P is less than that of point Q. The x coordinates of Q and P are, respectively, $g^A(s_2, c_2) \equiv f^A(s_2, c_2)/s_2$ and $g^A(s_{2^*}, c_2) \equiv f^A(s_{2^*}, c_2)/s_{2^*}$. As W_y is a shock, we have $s_2 < s_{2^*}$. Since $f^A(s, c)$ is convex, it follows that $g^A(s_2, c_2) < g^A(s_{2^*}, c_2)$. Thus the components μ and ν of the normal vector are both positive. □

LEMMA 4.2. *For solutions in which W_y is an s-family rarefaction, $-\pi/2 < \theta_{\mu\nu} < 0$.*

Proof. Since W_y is an s-family rarefaction wave, Q in Figure 4.2(b) corresponds to a region over which the contact discontinuity C_x interacts with W_y. Figure 4.3(b) illustrates the details of the construction of this interaction. As W_y is a rarefaction, $s_2 > s_{2^*}$ and $g^A(s_2, c_2) > g^A(s_{2^*}, c_2)$. Given that the $(g^A(s, c_2), g^B(s, c_2))$ contact base coordinate curve is monotonic increasing, concave down, it is clear from the construction in Figure 4.3(b) that the normal to the straight segment of the contact discontinuity separating s_{2^*}, c_2 from s_e, c_1 obeys $-\pi/2 < \theta_{\mu\nu} < 0$. □

Figures 4.3(a) and 4.3(b) also indicate how the saturation values s_{1^*}, s_e and, in the CRCW$_x$ cases, the saturation values in the rarefaction fan W_{QR} are found as crossing points of the appropriate contact discontinuity tangent line with the $(g^A(s, c_1), g^B(s, c_1))$ contact base coordinate curve. (Figures 4.3(a) and 4.3(b) demonstrate this for the case $c_1 > c_2$. For $c_1 < c_2$ the computation is the same.) Lemma 4.3 addresses the computation of these saturation values.

LEMMA 4.3. *s_{1^*} is always the smaller root of the quadratic equation*

$$A_1(s_{1^*})^2 - (1 + A_1)s_{1^*} + g_2^A = 0, \tag{4.1}$$

where $g_2^A \equiv g^A(s_2, c_2)$ and $A_1 \equiv A(1 - c_1)$. s_{2^*} and s_e are computed analogously.

Proof. For given values of s_2, c_1, and c_2, the saturation s_{1^*} is obtained using the Rankine–Hugoniot relation

$$g^A(s_{1^*}, c_1) = g^A(s_2, c_2). \tag{4.2}$$

We easily get the quadratic equation (4.1) from this relation. There are two solutions, s_+, s_- $(s_+ > s_-)$, for s_{1^*}. As $s_+ > 1$ and $s_- \in (0,1)$, the correct choice is $s_{1^*} = s_-$. s_{2^*} can be computed analogously from the Rankine–Hugoniot relation

$$g^B(s_{2^*}, c_2) = g^B(s_1, c_1) \tag{4.3}$$

for given values of s_1, c_1, and c_2.

To compute s_e, let $\hat{n} = (\mu, \nu)$ denote the normal direction to the contact discontinuity C_{PQ}. Given s_{2^*}, c_1, and c_2, the value of s_e is determined from the Rankine–Hugoniot relation

$$g^{\hat{n}}(s_e, c_1) = g^{\hat{n}}(s_{2^*}, c_2), \tag{4.4}$$

where $g^{\hat{n}} = \mu g^A + \nu g^B$. Again (4.4) produces a quadratic equation for s_e with two solutions s_+, s_-. As $s_+ \in (-\infty, 0) \cup (1, \infty)$ and $s_- \in (0,1)$, the correct solution is $s_e = s_-$. ∎

The uniqueness of these saturation values follows from the following remark.

Remark 4.1. The tangent lines of the contact discontinuity C_{PQ} in Figure 4.3 cutting the c_2 contact base coordinate curve also cut the c_1 contact base coordinate curve exactly once. This follows immediately from Lemma 4.3, since there is always a unique $s \in (0,1)$ which satisfies the Rankine–Hugoniot relation (4.4).

Interaction at *P*. The s-family wave W_{PR} is required to provide the transition from the "intermediate state" s_e, c_1 to the initial data state s_1, c_1. We now discuss the determination of this wave type. The wave type depends on the position of the point P, which depends on the relative magnitudes of s_e and s_1, which (Proposition 4.5 below), depends only on the relative magnitudes of c_1 and c_2. We develop first a preparatory lemma needed to prove Proposition 4.5.

Let s_0, c_0 be a fixed state value, $\hat{n} = (\mu, \nu)$ be a unit vector, and $\theta_{\mu\nu} = \tan^{-1}(\nu/\mu)$ as before. Consider the curve in phase space defined by the condition

$$g^{\hat{n}}(s, c) = g^{\hat{n}}(s_0, c_0). \tag{4.5}$$

Using the form for $g^\alpha, \alpha = A, B$, specific to our model, the curve defined by (4.5) can be rewritten explicitly as a function of s,

$$c_0^{\mu,\nu}(s) \equiv \frac{s(\mu + \nu + (\mu A + \nu B)(1-s)) - (\mu g_0^A + \nu g_0^B)}{(\mu A + \nu B) s(1-s)}. \tag{4.6}$$

Note that $c_1^{1,0}(s)$ and $c_1^{0,1}(s)$ correspond, respectively, to the terse notation $c_1^A(s)$ and $c_1^B(s)$ introduced earlier. For our model it is easy to show that $c_0^\alpha(s)$, $\alpha = A, B$, is monotonic increasing on $s \in (0,1)$ for any constant c_0. As noted in Figure 4.2(a), the curves $c_1^A(s)$ and $c_1^B(s)$ divide phase space into the four regions displaying different behaviors at infinity. We consider the relationship between the curves $c_0^A(s)$, $c_0^{\mu,\nu}(s)$, and $c_0^B(s)$.

LEMMA 4.4.

1. *If* $0 < \theta_{\mu\nu} < \pi/2$, *then* $c_0^A(s) > c_0^{\mu,\nu}(s) > c_0^B(s)$ *when* $s > s_0$; $c_0^A(s) < c_0^{\mu,\nu}(s) < c_0^B(s)$ *when* $s < s_0$.

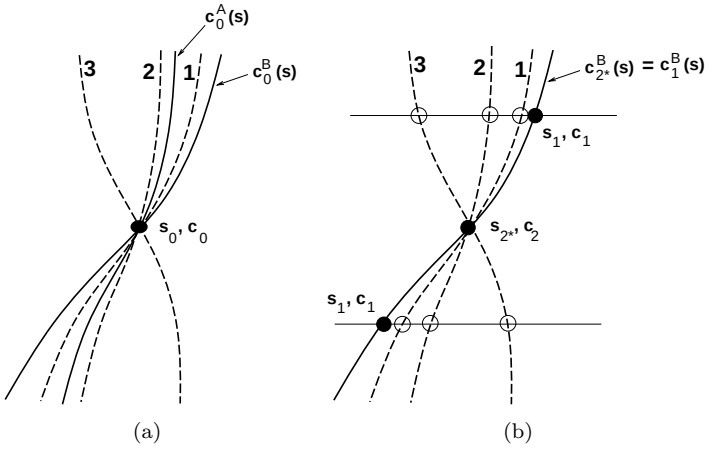


FIG. 4.4. (a) *The behavior of the curve* $c_0^{\mu,\nu}(s)$ *(dashed curves) relative to* $c_0^A(s)$ *and* $c_0^B(s)$ *for the three cases of Lemma* 4.4. (b) *Behavior of the location of the state* s_e, c_1 *(open points) for* $c_2 > c_1$ *and* $c_2 < c_1$.

2. *If* $-\pi/2 < \theta_{\mu\nu} < 0$ *and* $\mu A + \nu B > 0$, *then* $c_0^{\mu,\nu}(s) > c_0^A(s) > c_0^B(s)$ *when* $s > s_0$; $c_0^B(s) > c_0^A(s) > c_0^{\mu,\nu}(s)$ *when* $s < s_0$.
3. *If* $-\pi/2 < \theta_{\mu\nu} < 0$ *and* $\mu A + \nu B < 0$, *then* $c_0^{\mu,\nu}(s) > c_0^B(s) > c_0^A(s)$ *when* $s < s_0$; $c_0^A(s) > c_0^B(s) > c_0^{\mu,\nu}(s)$ *when* $s > s_0$.

Proof. These conclusions follow easily from examination of the equations

$$c_0^B(s) - c_0^{\mu,\nu}(s) = \frac{(A-B)\mu}{Bs(1-s)(\mu A + \nu B)}(s - s_0), \tag{4.7}$$

$$c_0^{\mu,\nu}(s) - c_0^A(s) = \frac{(B-A)\nu}{As(1-s)(\mu A + \nu B)}(s - s_0), \tag{4.8}$$

and noting $0 < A < B < 1/2$, $s \in (0,1)$. ∎

The results of Lemma 4.4 are summarized in Figure 4.4(a).

PROPOSITION 4.5. *If* $c_2 > c_1$, *then* $s_e > s_1$; *if* $c_2 < c_1$, *then* $s_e < s_1$.

Proof. The proof follows from the application of Lemma 4.4 with $s_0, c_0 = s_{2^*}, c_2$. Construction of the contact discontinuity C_y requires $c_1^B(s) = c_{s_2^*}^B(s)$ as shown in Figure 4.4(b). Thus if $c_2 > c_1$, all possible locations for the state s_e, c_1 lie to the right of s_1, c_1; if $c_2 < c_1$, all possible locations for the state s_e, c_1 lie to the left of s_1, c_1. ∎

We now address the nature of the W_{PR} wave.

PROPOSITION 4.6. *For* $c_2 > c_1$, W_{PR} *is an* S^+ *shock.*

Proof. The proof relies on the relationship between the shock base curves and the contact base coordinate lines. Since $c_2 > c_1$, by Proposition 4.5 $s_e > s_1$. Consider Figure 4.5 which shows the rarefaction base points $(f_s(s_1; c_1),\ f_s(s_e; c_1))$, the shock base point $\sigma(s_e, s_1; c_1)$, and the shock base curve $\sigma(s, s_1; c_1)$. This shock base curve ends at the contact base point $(g^A(s_1, c_1), g^B(s_1, c_1))$ which lies on the $c = c_1$ contact base coordinate curve. The c-family contact discontinuity separating the states s_1, c_1 and s_{2^*}, c_2 is a horizontal line (since this Riemann problem is "at infinity") and must pass through the contact base points $(g^A(s_1, c_1), g^B(s_1, c_1))$ and $(g^A(s_{2^*}, c_2), g^B(s_{2^*}, c_2))$. Since $c_2 > c_1$ the $c = c_2$ contact base coordinate curve

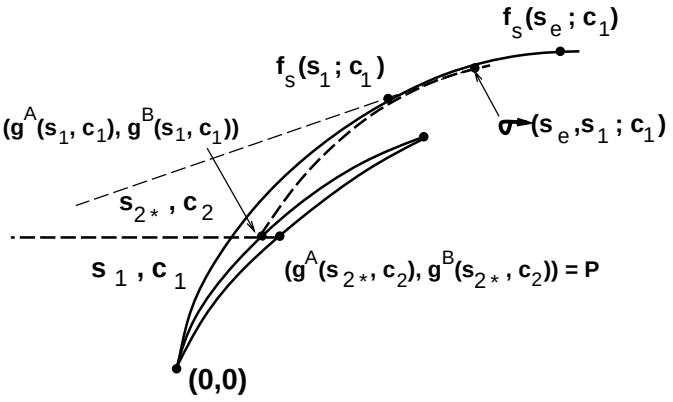


FIG. 4.5. *For $c_2 > c_1$ (i.e., $s_e > s_1$) the contact base point $(g^A(s_{2^*}, c_2), g^B(s_{2^*}, c_2))$ which is the point P in Figure* 4.2(b) *must lie below and to the right of the contact base point $(g^A(s_1, c_1), g^B(s_1, c_1))$. Consequently, W_{PR} must be an S^+ shock.*

must lie to the right of the $c = c_1$ contact base coordinate curve. Hence the contact base point $(g^A(s_{2^*}, c_2), g^B(s_{2^*}, c_2))$ must lie at the intersection of the horizontal line through $(g^A(s_1, c_1), g^B(s_1, c_1))$ and the $c = c_2$ contact base coordinate curve. However, $(g^A(s_{2^*}, c_2), g^B(s_{2^*}, c_2))$ is just the point P in Figure 4.2(b). Part of the boundary of the $S^+_{s_e,s_1}$ region is formed by the half line beginning at $\sigma(s_e, s_1; c_1)$ and passing through the rarefaction base point $f_s(s_1; c_1)$ (light dashed line in Figure 4.5). The relation (see section 2.4) between the rarefaction $f_s(s; c_1)$ and shock base curve $\sigma(s, s_1; c_1)$ guarantees that $(g^A(s_1, c_1), g^B(s_1, c_1))$ lies to the right of this boundary. Consequently, P also lies to the right of this boundary, i.e., P lies within the $S^+_{s_e,s_1}$ region. □

PROPOSITION 4.7. *For $c_2 < c_1$, W_{PR} is either* (i) *an R^+ rarefaction,* (ii) *an S^-R^+ composite, or* (iii) *an S^- shock.*

Proof. As $c_2 < c_1$, by Proposition 4.5 $s_e < s_1$. The resolution of the discontinuity between the states s_e, c_1 and s_1, c_1 involves Figure 3.1(b) where s_1, s_e and c_1 correspond, respectively, to the values s_l, s_r and c in Figure 3.1(b). If $g^B(s_{2^*}, c_2)$ (which is the y coordinate of point P) is less than $f^B_s(s_e, c_1)$ (which is the y coordinate of the rarefaction base point denoted $f_s(s_r; c)$ in Figure 3.1(b)), clearly point P can lie only in the R^+, S^-R^+, or S^- regions indicated in the figure.

We need therefore consider only the case $f^B_s(s_e, c_1) < g^B(s_{2^*}, c_2)$. Since $c_2 < c_1$, the contact base coordinate curve $g(s; c_2)$ lies above the curve $g(s; c_1)$. The contact discontinuity C_{PQ} crosses the contact base coordinate curve $g(s; c_2)$ at the base point $P = (g^A(s_{2^*}, c_2), g^B(s_{2^*}, c_2))$ and the curve $g(s; c_1)$ at the contact base point $(g^A(s_e, c_1), g^B(s_e, c_1))$. The horizontal contact C_y passes through the base points P and $(g^A(s_1, c_1), g^B(s_1, c_1))$. Since the tangent to $g(s; c_1)$ at the base point $(g^A(s_e, c_1), g^B(s_e, c_1))$ has to meet the rarefaction base curve $f_s(s; c_1)$ at the base point $(f^A_s(s_e, c_1), f^B_s(s_e, c_1))$, and since both the rarefaction and contact base coordinate curves are concave down, this rarefaction base point $(f^A_s(s_e, c_1), f^B_s(s_e, c_1))$ has to be on a segment of the rarefaction base curve $f_s(s; c_1)$ lying inside the "triangular" region having vertices P, $(g^A(s_e, c_1), g^B(s_e, c_1))$ and $(g^A(s_1, c_1), g^B(s_1, c_1))$ as shown in Figure 4.6. (Note that we are considering the case $f^B_s(s_e, c_1) < g^B(s_{2^*}, c_2)$.) Since the rarefaction base curve is concave down, point P is located left of the tangent line to the rarefaction base curve $f_s(s; c_1)$ at the point $(f^A_s(s_e, c_1), f^B_s(s_e, c_1))$.

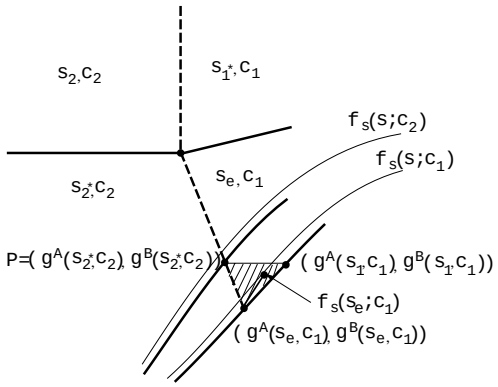


FIG. 4.6. *For $c_2 < c_1$, the rarefaction base point $(f_s^A(s_e, c_1), f_s^B(s_e, c_1))$ has to be on a segment of the rarefaction base curve $f_s(s; c_1)$ lying inside the "triangular" region if $f_s^B(s_e, c_1) < g^B(s_{2*}, c_2)$. This figure is for the $CSCW_x$ case; the $CRCW_x$ case has a similar "triangular" region.*

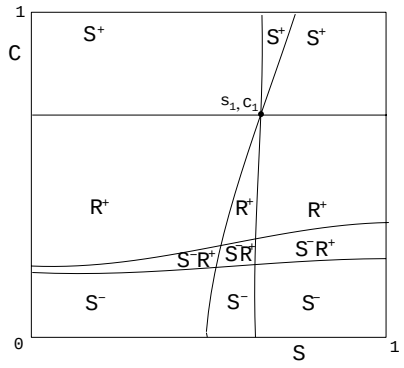


FIG. 4.7. *Identity of the s-family wave W_{PR} as a function of the location in phase space of (s_2, c_2) relative to (s_1, c_1).*

This implies point P can lie only in the S^-R^+ or S^- regions indicated in Figure 3.1(b). □

Figure 4.7 summarizes the results for the identity of the s-family wave W_{PR} according to the location of the initial state s_2, c_2 relative to s_1, c_1 in phase space. There are 12 regions formed by the combined classification of the solution behavior at infinity and the W_{PR} wave type.

This classification is generic with the following exception. Due to the finite limits on the values of s and c in the model, as the value of c_1 is lowered, the S^- and S^-R^+ regions in Figure 4.7 will vanish. Additionally, the width of the S^-R^+ region varies with c_1, increasing in width as c_1 decreases. As $c_1 \to 1$, the S^-R^+, R^+, and S^+ regions shrink to zero area. These observations are based upon numerical computations designed to identify the phase space boundaries of the S^+, R^+, S^-R^+, and S^- regions. The numerical computations vary s_1, c_1 over their full range of values and for each fixed pair s_1, c_1 perform the classification for a extensive choice of values of s_2, c_2 covering phase space on a fine grid ($\Delta s = 2 \cdot 10^{-4}$, $\Delta c = 1 \cdot 10^{-4}$). An illustration of the results for six choices of s_1, c_1 ($s_1 = 0.59$ held fixed, and c_1 varying over the range [0.4,0.95]) is given in Figure 4.8.

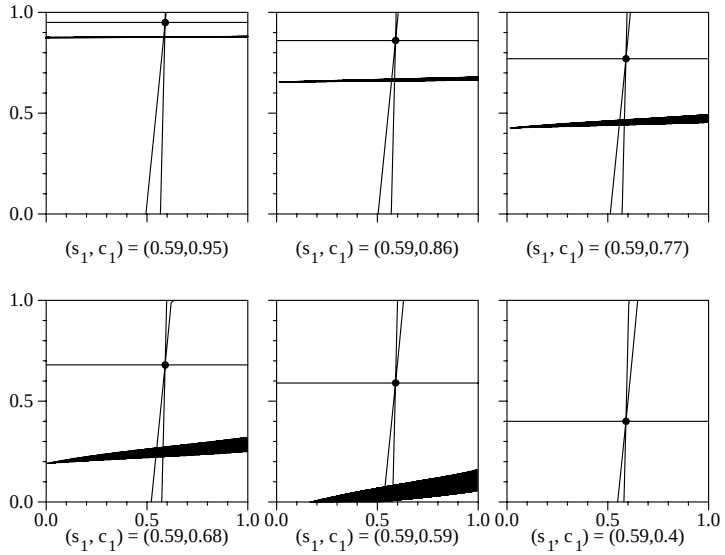


FIG. 4.8. *Due to the finite range of c and s, certain realizations of the wave type W_{PR} may not appear for all values of s_1, c_1.*

5. Solution construction for the single quadrant Riemann problem. In this section, we present the detailed solutions for each of the 12 solution types in Figure 4.7 and discuss the final wave interaction in region R for each case.

For each solution construction the same general steps are used. With reference to Figure 4.2(b), they are as follows:

- From separate 1-D Riemann problems, determine the states s_{1^*} and s_{2^*} (Lemma 4.3) and the s-family waves W_x, W_y "at infinity."
- The interaction at Q of the s-family wave W_y (separating state (s_2, c_2) and (s_{2^*}, c_2)) with the contact discontinuity C_x separating state (s_2, c_2) and (s_{1^*}, c_1) is resolved.
- The location of the point P of intersection of the contact discontinuities C_y and C_{PQ} is determined. It has coordinates $P = (g^A(s_{2^*}, c_2), g^B(s_{2^*}, c_2))$.
- The direction of the contact discontinuity C_{PQ} separating constant states s_{2^*}, c_2 and s_e, c_1 is determined.
- The value of the intermediate state saturation s_e is determined (Lemma 4.3).
- The s-family wave W_{QR} separating constant states (s_{1^*}, c_1) and (s_e, c_1) is resolved.
- The s-family wave W_{PR} separating constant states (s_1, c_1) and (s_e, c_1) is resolved.
- The interaction involving the s-family waves W_x, W_{QR}, and W_{PR} is resolved in the region R.

We present each solution for all 12 cases. However, due to space limitations, we display only simplified figures of each solution in Figure 5.1. For complete figures, we refer to [2]. We focus our discussion on the region R.

5.1. CSCS. The solution contains only contact discontinuities (dashed lines) and s-family shocks (solid). Each contact discontinuity is a straight line whose tangent passes through the two relevant contact base points. Each shock separates two constant states and is therefore a straight line whose tangent passes through the appropriate shock base point. W_y and W_{QR} are S^- shocks. (The shock-contact in-

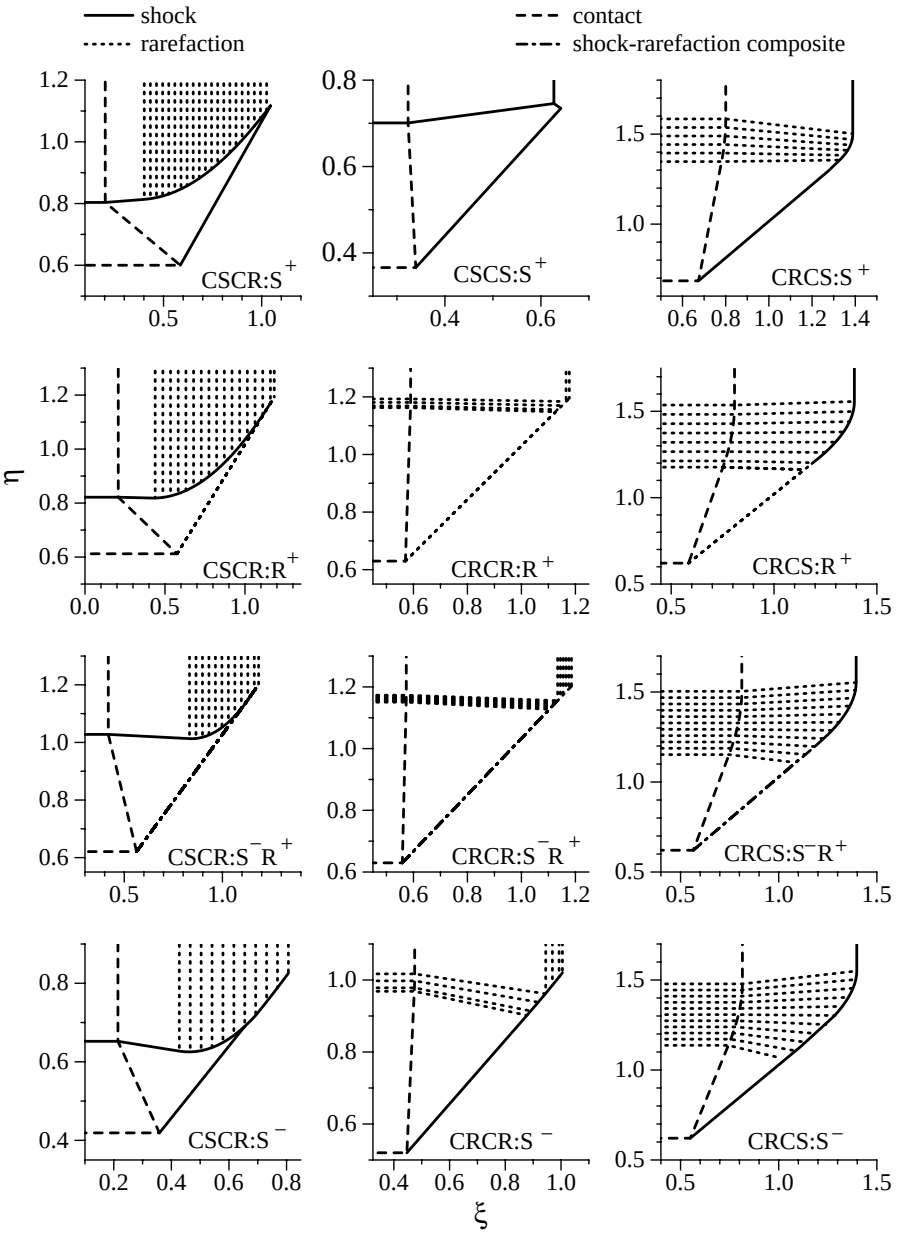


FIG. 5.1. *All* 12 *solutions in the* (ξ, η) *plane.*

teraction at the point Q preserves shock type.) The shock W_x and the short third shock emerging from the W_{QR}, W_x interaction point are also of type S^-. W_{PR} is a shock of type S^+ by Proposition 4.6.

5.2. CSCR. The CSCR case breaks into four subcases, one for $c_2 > c_1$ and three for $c_2 < c_1$.

5.2.1. $\boldsymbol{c_2 > c_1}$**.** As for the CSCS case, the waves W_y and W_{QR} are S^- type shocks; W_{PR} is always a shock of type S^+ by Proposition 4.6. W_x is a rarefaction

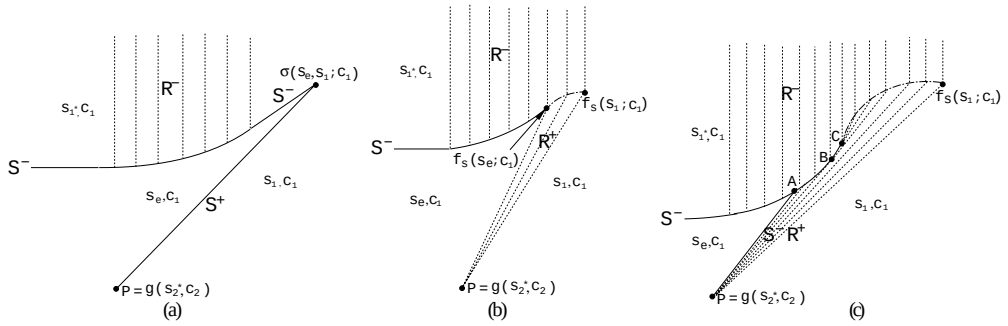


FIG. 5.2. *CSCR solution: Details of the W_{QR} and W_{PR} wave interaction at point R for CSCR solutions when the W_{PR} wave is of type* (a) S^+, (b) R^+, (c) S^-R^+.

fan of type R^-. Each characteristic wave in the R^- fan is a straight half line whose tangent passes through to the appropriate rarefaction base point. Details of the interaction are given in Figure 5.2(a). The interaction of the S^- shock W_{QR} and the R^- rarefaction fan results in a continuously curving shock of decaying strength. As $s_e > s_1$, the S^- shock interacts with the R^- fan and "emerges" from this interaction with a straight segment that ends at the shock base point $\sigma(s_e, s_1; c_1)$. The S^+ shock W_{PR} also terminates at this shock base point.

5.2.2. $c_2 < c_1$. There are three different cases, labeled by the possible type for the W_{PR} wave.

$\mathbf{R}^+$. For a range of values of c_2, W_{PR} is an R^+ rarefaction fan centered at the point P. (P is the contact base point $(g^A(s_{2^*}, c_2), g^B(s_{2^*}, c_2))$.) W_y and W_{QR} are S^- shocks, and W_x is an R^- rarefaction fan. Details of the W_x, W_{QR} and W_{PR} interaction are given in Figure 5.2(b). Upon interacting with the R^- rarefaction, the S^- shock begins to curve and weaken. One edge of the R^+ rarefaction fan centered at P also meets the S^- shock at this base point. Over the range of state values (s_e, c_1) to (s_1, c_1) the characteristic lines of both R^+ and R^- rarefaction fans meet along the section of the rarefaction base curve between the base points $f_s(s_e; c_1)$ and $f_s(s_1; c_1)$.

This $\mathrm{W}_{QR} \leftrightarrow \mathrm{W}_x$ wave interaction is of the type $S^- \leftrightarrow R^-$. Theorem 1 of Zhang and Zhang [20], based upon analysis of the four quadrant Riemann problem for a single conservation law, states that either "the S^- shock will penetrate the R^- wave completely" or "during the penetration, an R^+ wave will form having the S^- shock as an envelope of characteristic lines of the R^+ (i.e., an S^-R^+ composite wave is formed)." In our entropy obeying solution, however, the S^- shock terminates with zero strength at the rarefaction base point $f_s(s_e; c_1)$ and neither "complete penetration" nor "composite wave formation" occurs. It is currently unclear as to whether the presence of the W_{PR} R^+ wave emanating from P is the reason for the behavioral difference between our observed solution and the Zhang–Zhang theorem. (We do, however, find composite waves occurring in $S^- \leftrightarrow R^-$ interactions in our S^-R^+ labeled subcases of CSCR, CRCS, and CRCR.)

$\mathbf{S}^-\mathbf{R}^+$. For a range of values of c_2, W_{PR} becomes an S^-R^+ composite wave with the R^+ fan centered at the point P (the contact base point $(g^A(s_{2^*}, c_2), g^B(s_{2^*}, c_2))$. The S^-R^+ composite wave W_{PR} is very narrow; Figure 5.2(c) gives an enlarged sketch of this wave and the interaction region of the waves W_{QR}, W_x, and W_{PR}. The straight S^- shock begins to curve upon interaction with the upper R^- rarefaction

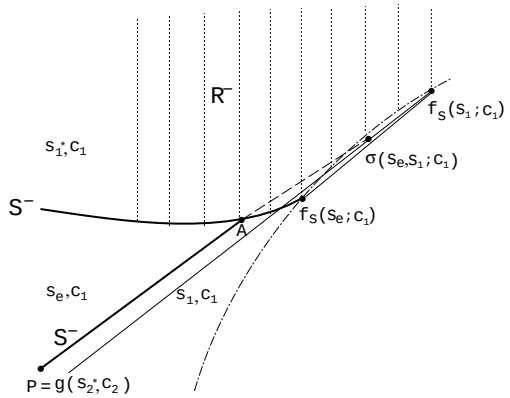


FIG. 5.3. *Details of the W_{QR} and W_{PR} wave interaction at point R for the CSCR solution with $c_2 < c_1$. (The W_{PQ} wave is of type S^-.)*

fan. At point A it also interacts with the S^- shock from the lower S^-R^+ composite wave. The resultant shock (curve segment AB) of S^- type separates the R^- and R^+ rarefaction fans. At point B a characteristic of the R^+ fan meets this S^- shock tangentially. On the segment AB the effective flux function $\hat{n} \cdot (f^A, f^B)$, where $\hat{n}$ is normal to the shock, remains convex. On the segment BC the effective flux develops a single inflection point and a composite S^-R^+ wave forms with the characteristics of this R^+ wave "emerging" tangentially from the S^- shock. This second R^+ fan "merges" continuously with the remainder of the R^+ fan centered on point P. At C the S^- shock decays to zero strength (the characteristic state (s^*, c_1) on both sides of the shock is the same) and meets the rarefaction base curve at the base point $f_s(s^*; c_1)$. Over the segment of rarefaction base curve from C to $f_s(s_1; c_1)$ the characteristics of the upper R^- and lower R^+ rarefaction waves meet continuously. Composite waves of the type generated over the segment BC have been seen in the work of Zhang and Zheng [22] and even as early as the work by Wagner [17]. In these earlier works, however, one side of the "precursor" shock segment (i.e., the shock segment AB) was always a constant state. In our case the state on both sides of the precursor segment is changing continuously.

S^-. For the final range of values of $c_2 < c_1$, W_{PR} is an S^- shock, W_{QR} is an S^- shock, and W_x is an R^- rarefaction.

The following lemma shows that the W_{QR} S^- shock must interact with the W_{PR} S^- shock before W_{QR} has "completed" its interaction with the W_x R^- rarefaction.

LEMMA 5.1. *The W_{PR} and W_{QR} S^- shock waves must interact when the state on the upper side of W_{PR} has a saturation value lying in the range (s_{1^*}, s_e).*

Proof. The proof is geometrical. Figure 5.3 shows the rarefaction base curve, the base points $f_s(s_1; c_1)$, $f_s(s_e; c_1)$, and the shock base point $\sigma(s_1, s_e; c_1)$ which lies in their convex hull. As W_{PR} is an S^- shock, the point P must lie to the left of the straight dashed line which passes through the two base points $f_s(s_1; c_1)$ and $\sigma(s_1, s_e; c_1)$. Assume the S^- shock W_{QR} *does not* interact with W_{PR} but only with the R^- rarefaction fan W_x. Then, as shown in Figure 5.3, the shock W_{QR} must end at zero strength at the rarefaction base point $f_s(s_e; c_1)$. However, we see that this is geometrically impossible without an interaction with the W_{PR} shock. Hence the two shocks W_{PR} and W_{QR} must interact as claimed. □

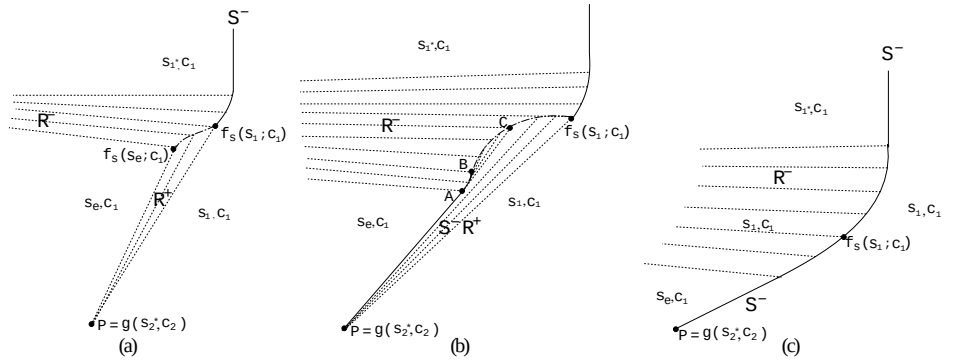


FIG. 5.4. *Details of the W_{QR} and W_{PR} wave interaction at point R for CRCS solutions when the W_{PR} wave is of type* (a) R^+, (b) S^-R^+, (c) S^-.

The shocks W_{PR} and W_{QR} must interact at a point A. From A to $f_s(s_1; c_1)$ an S^- shock separates the R^- rarefaction from the constant state (s_1, c_1). This shock ends with zero strength at the base point $f_s(s_1; c_1)$.

5.3. CRCS. The CRCS case also breaks into four subcases, one for $c_2 > c_1$ and three for $c_2 < c_1$.

5.3.1. $c_2 > c_1$. W_y and W_{QR} are R^- rarefaction fans; W_{PR} is an S^+ shock by Proposition 4.6; W_x is an S^- shock. We note that, as in the isotropic case [3], the contact-rarefaction interaction produces a dynamic diffraction of the W_{QR} rarefaction fan, *in this case focusing the fan.* The S^- shock interacts with the W_{QR} fan; both S^+ and S^- shocks meet at the shock base point $\sigma(s_1, s_e; c_1)$.

5.3.2. $c_2 < c_1$. There are three different cases, again labeled by the possible type for the W_{PR} wave. In all cases the W_y and W_{QR} waves are rarefactions; interaction with the C_x contact discontinuity produces diffraction of the W_{QR} fan.

$\mathbf{R}^+$. For a range of values of $c_2 < c_1$, a centered R^+ rarefaction fan forms at the point P. Details of the W_x, W_{QR} and W_{PR} interaction are given in Figure 5.4(a). The saturations s_e, s_1 and s_{1^*} are ordered $s_e < s_1 < s_{1^*}$. W_x is an S^- shock. Due to the relative ordering of s_e, s_1 and s_{1^*}, W_x interacts with W_{QR} until it decays to zero strength at the rarefaction base point $f_s(s_1; c_1)$. The remainder of the W_{QR} rarefaction and W_{PR} meet continuously along the rarefaction base curve between the points $f_s(s_1; c_1)$ and $f_s(s_e; c_1)$.

$\mathbf{S}^-\mathbf{R}^+$. For the next lowest range of values of $c_2 < c_1$, the W_{PR} wave becomes an S^-R^+ composite wave with the R^+ fan centered at the point P. Details of the interaction is given in Figure 5.4(b). W_x interacts with the W_{QR} fan and terminates tangentially, at zero strength, at the base point $f_s(s_1; c_1)$. The S^- shock in the composite W_{PR} wave interacts with the W_{QR} fan at A. It curves as a result of the interaction; over a segment (AB in Figure 5.4(b)) the effective flux function $\hat{n}\cdot(f^A, f^B)$ continues to have an inflection point and a composite S^-R^+ wave continues to form with the characteristics of the R^+ fan "emerging" tangentially from the shock along the segment AB. This fan merges continuously with the R^+ fan centered on point P. The point B is a rarefaction base point $f_s(\bar{s}; c_1)$ for some saturation $\bar{s}$ in the interval (s_e, s_1). Between the respective characteristics $(\bar{s}, c_1)$ and (s_1, c_1) the $\mathrm{W}_{QR}R^-$ and the R^+ rarefaction fans meet continuously along the rarefaction base curve.

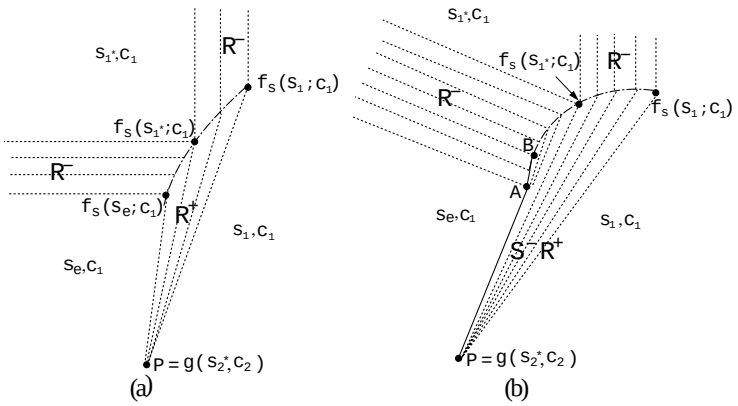


FIG. 5.5. *Details of the W_{QR} and W_{PR} wave interaction at point R for the CRCR solutions when the W_{PR} wave is of type* (a) R^+, (b) S^-R^+.

$\mathbf{S^-}$. For the lowest range of values of c_2, W_{PR} is an S^- shock. The S^- shocks W_x and W_{PR} interact separately with either end of the R^- rarefaction W_{QR}. Both shocks meet tangentially at the rarefaction base point $f_s(s_1; c_1)$, decaying to zero strength at this point. Details of the interaction are given in Figure 5.4(c).

5.4. CRCR. In this case, $c_2 < c_1$. We have the state ordering $s_e < s_{1^*} < s_1$. W_y, W_{QR}, and W_x are all R^- rarefaction fans. Again there are three possibilities for W_{PR}.

$\mathbf{R^+}$. For a range of values of $c_2 < c_1$, a centered R^+ rarefaction fan forms at the point P. Details of the W_x, W_{QR} and W_{PR} interaction are given in Figure 5.5(a). All rarefaction fans meet continuously along the rarefaction base curve between the base points $f_s(s_e; c_1)$ and $f_s(s_1; c_1)$. There are no s-family shocks in the solution.

$\mathbf{S^-R^+}$. For the next range of values of $c_2 < c_1$, W_{PR} forms an S^-R^+ composite wave with the R^+ fan centered at the point P. The W_{PR} R^+ rarefaction meets the upper part of the W_{QR} R^- rarefaction continuously along the rarefaction base curve between the base points $f_s(s_{1^*}; c_1)$ and $f_s(s_1; c_1)$. The interaction of the composite wave W_{PR} and the lower part of the W_{QR} R^- rarefaction is given in Figure 5.5(b) and is the same as in Figure 5.4(b).

$\mathbf{S^-}$. For the lowest range of values of $c_2 < c_1$, W_{PR} forms an S^- shock. It interacts separately with the W_x and W_{QR} rarefaction fans, meeting the base curve tangentially at the base point $f_s(s_1; c_1)$ where it decays to zero strength.

6. Discussion. It is tempting to comment on the generalizability of the construction technique utilized here to Riemann problems involving systems in $\mathbf{R}^2$. In $\mathbf{R}^2$ the integral curves (characteristic lines) for the ith family are given by (see, e.g., (2.12) of this paper, or (2.5) and following in [12])

$$\frac{\partial \eta(\xi)}{\partial \xi} = \lambda_i = \frac{\eta - R_{i,\eta}(U)}{\xi - R_{i,\xi}(U)}. \tag{6.1}$$

Here $U = (u_1, \ldots, u_n)$ denotes the solution. The rarefaction base points associated with this family are the points $(R_{i,\xi}(U), R_{i,\eta}(U))$. An ith family integral curve having state value U_0 must have a tangent line passing through the rarefaction base point $(R_{i,\xi}(U_0), R_{i,\eta}(U_0))$.

If the jth family is a contact discontinuity, the contact base points are similarly defined:

$$\frac{\partial \eta(\xi)}{\partial \xi} = \lambda_j = \frac{\eta - C_{j,\eta}(U)}{\xi - C_{j,\xi}(U)}; \tag{6.2}$$

the contact base points associated with this family are the points $(C_{j,\xi}(U), C_{j,\eta}(U))$. Such a contact discontinuity separating two states U_l and U_r must satisfy $\lambda_j(U_l) = \lambda_j(U_r)$; the tangent line to the contact discontinuity must pass through the two contact base points $(C_{j,\xi}(U_l), C_{j,\eta}(U_l))$ and $(C_{j,\xi}(U_r), C_{j,\eta}(U_r))$.

Finally a (smooth) discontinuity for the ith family satisfies a Rankine–Hugoniot condition

$$\frac{\partial \eta(\xi)}{\partial \xi} = \sigma_i = \frac{\eta - S_{i,\eta}(U_l, U_r)}{\xi - S_{i,\xi}(U_l, U_r)}. \tag{6.3}$$

(See, e.g., (2.16) of this paper, or (2.8.2) in [12].) The shock base points associated with this family are the points $(S_{i,\xi}(U_l, U_r), S_{i,\eta}(U_l, U_r))$; a shock discontinuity separating states U_l and U_r must have a tangent line passing through this shock base point.

Base points provide ODE integration (tangent) directions for the characteristic curves and smooth discontinuity curves, enabling the tracing of such curves in the ξ, η plane. Constant state conditions on characteristic curves or changing state conditions on each side of discontinuity curves implicitly define a (presumably) continuous path through a relevant set of base points. In this paper, the second state variable $u_2 = c$ is a Riemann invariant for the s-family of waves, and the rarefaction base points form curves (parametrized by values of c) vastly simplifying characteristic curve tracing. A simplifying organization of the set of shock base points also follows from the same invariance.

It is, however, the development of relationships between base points of different types (i.e., between the functions R_i, C_i, and S_i above) that seems to be critical in achieving "analytic" solution. The critical question is whether one can find general relationships that hold independent of the model under consideration. It is encouraging that equations such as (2.26) and (2.27) between the R and S functions for the same wave family are somewhat general relationships which hold in any scalar region of the solution and are independent of the form of the flux function f. (By scalar region we mean any region in which the variables associated with all but one wave family are constant.)

REFERENCES

[1] J. Guckenheimer, *Shocks and rarefactions in two space dimensions*, Arch. Rational Mech. Anal., 59 (1975), pp. 281–291.

[2] W. Hwang, *A Study of the* 2*-Dimensional Riemann Problem for a* 2×2 *Hyperbolic Conservation Law*, Ph.D. thesis, SUNY at Stony Brook, Stony Brook, NY, 2000.

[3] W. Hwang and W. B. Lindquist, *The* 2*-dimensional Riemann problem for a* 2×2 *hyperbolic conservation law* I. *Isotropic media*, SIAM J. Math. Anal., 34 (2002), pp. 341–358.

[4] E. Isaacson, *Global Solution of a Riemann Problem for a Non-Strictly Hyperbolic System of Conservation Laws Arising in Enhanced Oil Recovery*, preprint, The Rockefeller University, New York, NY, 1980.

[5] B. Keyfitz and H. Kranzer, *A system of non-strictly hyperbolic conservation laws arising in elasticity theory*, Arch. Rational Mech. Anal., 72 (1980), pp. 219–241.

[6] S. N. Kruzkov, *First order quasilinear equations in several independent variables*, J. Mat. USSR-Sb., 10 (1970), pp. 217–243.

[7] P. D. Lax and X. D. Liu, *Solution of two-dimensional Riemann problems of gas dynamics by positive schemes*, SIAM J. Sci. Comput., 19 (1998), pp. 319–340.
[8] W. B. Lindquist, *The scalar Riemann problem in two spatial dimensions: Piecewise smoothness of solutions and its breakdown*, SIAM J. Math. Anal., 17 (1986), pp. 1178–1197.
[9] W. B. Lindquist, *Construction of solutions for two-dimensional Riemann problems*, Comput. Math. Appl. Part A, 12 (1986), pp. 615–630.
[10] C. W. Schulz-Rinne, *Classification of the Riemann problem for two-dimensional gas dynamics*, SIAM J. Math. Anal., 24 (1993), pp. 76–88.
[11] C. W. Schulz-Rinne, J. P. Collins, and H. M. Glaz, *Numerical solution of the Riemann problem for two-dimensional gas dynamics*, SIAM J. Sci. Comput., 14 (1993), pp. 1394–1414.
[12] D. Tan and T. Zhang, *Two-dimensional Riemann problem for a hyperbolic system of nonlinear conservation laws* (I): *Four-J cases*, J. Differential Equations, 111 (1994), pp. 203–254.
[13] D. Tan and T. Zhang, *Two-dimensional Riemann problem for a hyperbolic system of nonlinear conservation laws* (II): *Initial data consists of some rarefaction*, J. Differential Equations, 111 (1994), pp. 255–283.
[14] D. Tan, T. Zhang, and Y. Zheng, *Delta-shock waves as limits of vanishing viscosity for hyperbolic system of conservation laws*, J. Differential Equations, 112 (1994), pp. 1–32.
[15] B. Temple, *Global solution of the Cauchy problem for a class of* 2×2 *nonstrictly hyperbolic conservation laws*, Adv. in Appl. Math., 3 (1982), pp. 335–375.
[16] A. I. Vol'pert, *The spaces BV and quasilinear equations*, J. Mat. USSR-Sb., 2 (1967), pp. 225–267.
[17] D. H. Wagner, *The Riemann problem in two space dimensions for a single conservation law*, SIAM J. Math. Anal., 14 (1983), pp. 534–559.
[18] S. Yang and T. Zhang, *The MmB difference solutions to the Riemann problem for a* 2-D *hyperbolic system of nonlinear conservation laws*, Impact Comput. Sci. Engrg., 3 (1991), pp. 146–180.
[19] P. Zhang, J. Li, and T. Zhang, *On two-dimensional Riemann problem for pressure-gradient equations of the Euler system*, Discrete Contin. Dynam. Systems, 4 (1998), pp. 609–634.
[20] P. Zhang and T. Zhang, *Generalized characteristic analysis and Guckenheimer structure*, J. Differential Equations, 152 (1999), pp. 409–430.
[21] T. Zhang and G. Chen, *Some fundamental concepts about systems of two spatial dimensions conservation laws*, Acta Math. Sinica, 6 (1986), pp. 463–474.
[22] T. Zhang and Y. Zheng, *Two dimensional Riemann problem for a single conservation law*, Trans. Amer. Math. Soc., 312 (1989), pp. 589–619.
[23] T. Zhang and Y. Zheng, *Conjecture on structure of solution of Riemann problem for* 2-D *gas dynamic systems*, SIAM J. Math. Anal., 21 (1990), pp. 593–630.
[24] T. Zhang and Y. Zheng, *Exact spiral solutions of the two dimensional compressible Euler equations*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 117–133.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 385–408

ERROR OF THE NETWORK APPROXIMATION FOR DENSELY PACKED COMPOSITES WITH IRREGULAR GEOMETRY*

LEONID BERLYAND† AND ALEXEI NOVIKOV‡

Abstract. We introduce a discrete network approximation to the problem of the effective conductivity of the high contrast, highly packed composites in which inclusions are irregularly (randomly) distributed in a hosting medium so that a significant fraction of them may not participate in the conducting spanning cluster. For this class of spacial arrays of inclusions we derive a discrete network approximation and obtain its a priori error estimate. We obtained an explicit dependence of the network approximation and its error on the irregular geometry of the inclusions' array. We use variational techniques to provide rigorous mathematical justification for the approximation and its error estimate.

Key words. effective conductivity, discrete network, error estimate, variational bounds

AMS subject classifications. 74Q05, 35Q72, 94C05

PII. S0036141001397144

1. Introduction.

We study the effective properties such as the effective conductivity or the effective dielectric constant of composite materials in which a large number of inclusions are irregularly (randomly) distributed in a homogeneous hosting medium (matrix). For ease of presentation and clarity we concentrate here on the effective conductivity. We are particularly interested in the case of the high contrast, highly packed particulate composites, that is, when the conductivity of the inclusions is much larger than the conductivity of the hosting medium and the volume fraction of the inclusions is very high. High contrast composites are extremely attractive for the design of new materials with physical properties better than those of their constituents. The case when the concentration of the filling inclusions is high is particularly relevant to polymer/ceramic composites, because a polymer matrix compensates for the brittle nature of ceramics which is their main weakness. A survey on the relevant engineering problems in two and three dimensions (fibers and particles in a matrix) can be found in [3].

Our main tool is (a modification of) the discrete network approximation (DNA) of [3] for a two-dimensional composite, where the inclusions are modeled as identical disks. We focus on the two key issues arising for this approximation. The first is the explicit error estimate of the DNA to the continuum problem of effective conductivity. The second is a quantitative estimate on how the connectivity patterns for various irregular distributions of the inclusions affect the effective conductivity.

The main advantage of our DNA is that it is easy to implement numerically and at the same time it captures geometric patterns of the location of inclusions in the matrix. The importance of the geometric patterns in evaluation of the effective properties of

*Received by the editors October 29, 2001; accepted for publication (in revised form) April 25, 2002; published electronically October 31, 2002. Part of this work was done while both authors were visiting members of MSRI, Berkeley.

http://www.siam.org/journals/sima/34-2/39714.html

†Department of Mathematics & Materials Research Institute, 414 McAllister Building, Pennsylvania State University, University Park, PA 16802 (berlyand@math.psu.edu). The work of this author was supported by NSF grant DMS-9971999.

‡IMA, University of Minnesota, 400 Lind Hall, 207 Church S.E., Minneapolis, MN 55455. Current address: Department of Applied & Computational Mathematics, California Institute of Technology, 1200 E. California Boulevard, MC 217-50, Pasadena, CA 91125 (novikov@acm.caltech.edu).

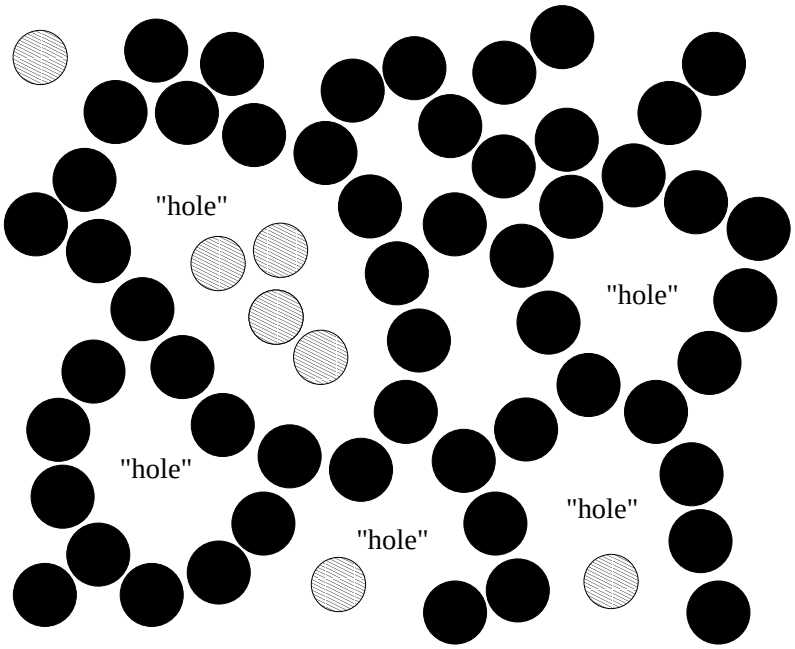
high contrast composites can be seen in the analysis of periodic structures. It was observed that for such periodic composites of moderate volume fraction, that is, away from the almost touching situation, the effective conductivity is of the order of the conductivity of the matrix (see, for example, [2], [16], [17], and references therein). In other words, the filler has almost no effect on the effective conductivity. However, in the case of almost touching inclusions, the effective conductivities of two periodic structures with different locations of inclusions in the matrix can be significantly different for the same volume fraction. For example, if the contrast ratio of the constituents is assumed to be ∞, then for the same volume fraction of disks (equal to $\pi/4$) for the hexagonal lattice, the effective conductivity $\hat{a} = O(1)$ (see [4]), while for the square lattice $\hat{a} = \infty$ (see [13]).

The case of irregularly distributed inclusions is not as well understood as the periodic case. Since the volume fraction of the inclusions is high, the irregular connectivity patterns in the whole composite (percolation effects) determine the behavior of the effective properties. Moreover, it was observed that the irregular connectivity patterns of conducting inclusions can greatly increase the effective conductivity. Therefore, there is a need for a simple model that is still able to capture percolation effects. Also, while for a given periodic structure the volume fraction of the inclusions uniquely determines the distances between the inclusions, this is no longer true for irregular structures, and one should search for a model with a new parameter which describes the local geometry when the inclusions are close to touching. Such a model (the network approximation) was proposed in [3] in two dimensions. The notion of the *interparticle distance parameter* for closely packed ("randomized" hexagonal) patterns, based on the Voronoi tessellation, was introduced there. In the present work we generalize this notion for a broad class of geometrical patterns. This is important in practical applications, because in real composites the array of the inclusions is often highly nonuniform due to the manufacturing process.

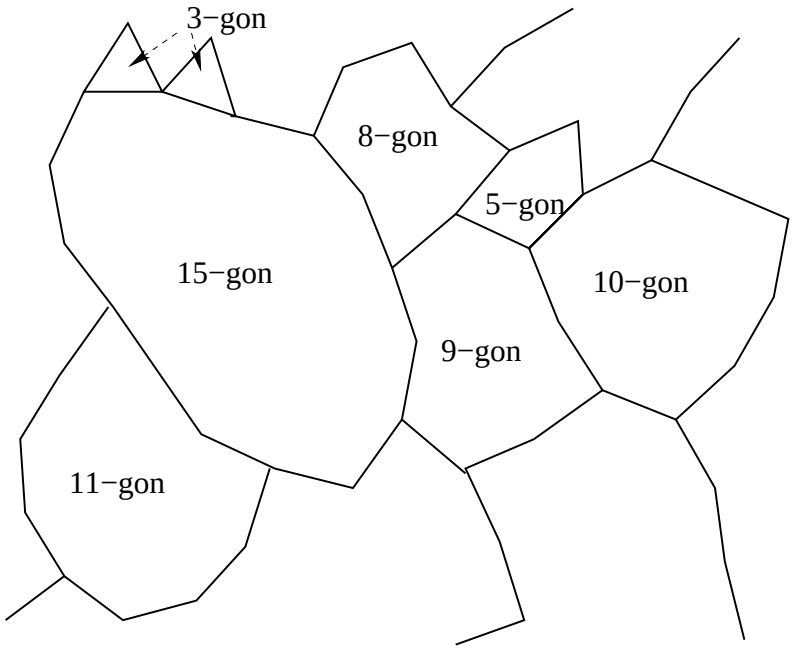
Our new approach allows us to derive an *explicit error estimate* for the DNA. Since most of the existing estimates provide an order of the magnitude of the error only, such explicit estimates are rare in homogenization theory. The class of geometrical patterns that can be handled by our approach includes a nonuniform irregular distribution, when a significant fraction of the inclusions does not participate in the conducting spanning cluster. This approach allows us to relax the close packing condition of [3] so that not all the "neighboring" inclusions (disks) are closely spaced. More specifically, we introduce and study the $\delta - \mathbf{N}$ close packing condition, which loosely speaking allows for "holes" with the perimeter of order $\mathbf{N}\,R$ in the conducting spanning cluster. Here R is the radius of the inclusions, and $\mathbf{N}$ is the number of inclusions in the perimeter of the largest hole in the conducting cluster (see Figure 1.1). Thus we account quantitatively for the presence of these holes in the composite.

The question of error estimates was raised by I. Babuska, because the analysis of [3] is asymptotic in the interparticle distance parameter and does not provide an error estimate. The analysis in the present paper does not use asymptotics, and it holds for any (small) finite value of the *relative* interparticle distance parameter. This enables us to prove the following error estimate for the effective conductivity $\hat{a}$:

$$\frac{|\hat{a} - I|}{I} \le C(\mathbf{N})\sqrt{\frac{\delta}{R}}, \tag{1.1}$$



(a) Black disks form the conducting cluster.
Hatched disks do not participate in the cluster.



(b) The graph that corresponds to the conducting cluster in (a). An **N**-gon $\mathbf{N} \geq 4$ corresponds to a "hole" of size **N**.

FIG. 1.1. *The conducting cluster in a composite with "holes".*

where I is the value of effective conductivity provided by the network approximation, δ/R is the relative interparticle distance, and for the constant $C(\mathbf{N})$ we provide an upper bound $C(\mathbf{N}) \leq 2.56\mathbf{N}^4$.

The discrete network models for various high contrast composites have been used extensively in the physics literature (see [1], [11], [12], [14], [18], [19]); however, the relation between the network and the underlying continuum problem was not studied there. In [15], high contrast conductivity problems were first formulated and analyzed using variational methods. There the high contrast field was of the form

$$e^{S(x)/\epsilon} \tag{1.2}$$

with a *smooth* function $S(x)$. In particular, the asymptotic analysis in the high contrast ratio parameter ϵ has been carried out in [15] for a random checkerboard model.

For the Kozlov's function (1.2) a *network* asymptotic approximation in the high contrast parameter ϵ was developed in [6], [7], [8], [9]. It was *rigorously proved* in [8] that the network approximates the original continuum problem. The analysis of [8] was carried out for high contrast continuum problems arising in imaging, when the materials' properties are not known and it is convenient to model the high contrast in a simple geometric manner by (1.2). In this model the key parameter, which determines the conductivity of the edges in the network, is $\sqrt{k_+/k_-}$, where k_+ and k_- are the principal curvatures at the saddle points of $S(x)$.

Our analysis applies to a class of physical problems where $S(x)$ is *not smooth*. In our case $S(x)$ is the characteristic function of the disks, $S(x) = 1$, $S(x) = 0$, on the inclusions and in the matrix, respectively (see Figure 2.1). Furthermore, in our case the high contrast parameter $\epsilon = 0$ and the analysis is carried out when the relative interparticle distance parameter is sufficiently small. In other words, we consider the infinite contrast material with ideally conducting inclusions. This assumption is valid for a variety of particulate composites (particles or fibers in a matrix), and it is in agreement with bounds [10] which imply that if the contrast ratio is greater than several hundred, then for practical purposes it can be taken to be infinite.

The paper is organized as follows. In section 2 we give the formulation of the problem and construct the modified DNA. In section 3 we formulate and prove the main result: an explicit analytical a priori error estimate for this approximation. We also present there numerical a posteriori error estimates.

2. Formulation.

2.1. Mathematical model. Consider a two-dimensional rectangular two-phase composite that consists of a matrix filled by a large number of inclusions. The inclusions are ideally conducting. Assume that all the inclusions are identical nonoverlapping disks. The centers of the disks are irregularly distributed in the rectangular domain. The distribution of the disks is dense; that is, the characteristic distance between two neighbors is much smaller compared to the radius of the disks.

Denote the domain occupied by the composite by $\Pi = [-L, L] \times [-1, 1]$ (Figure 2.1). Denote the disks that model the inclusions by D_i, $i = 1, \ldots, N$, where N is the total number of disks. Then

$$Q_p = \Pi \backslash \cup_{i=1}^{N} D_i \tag{2.1}$$

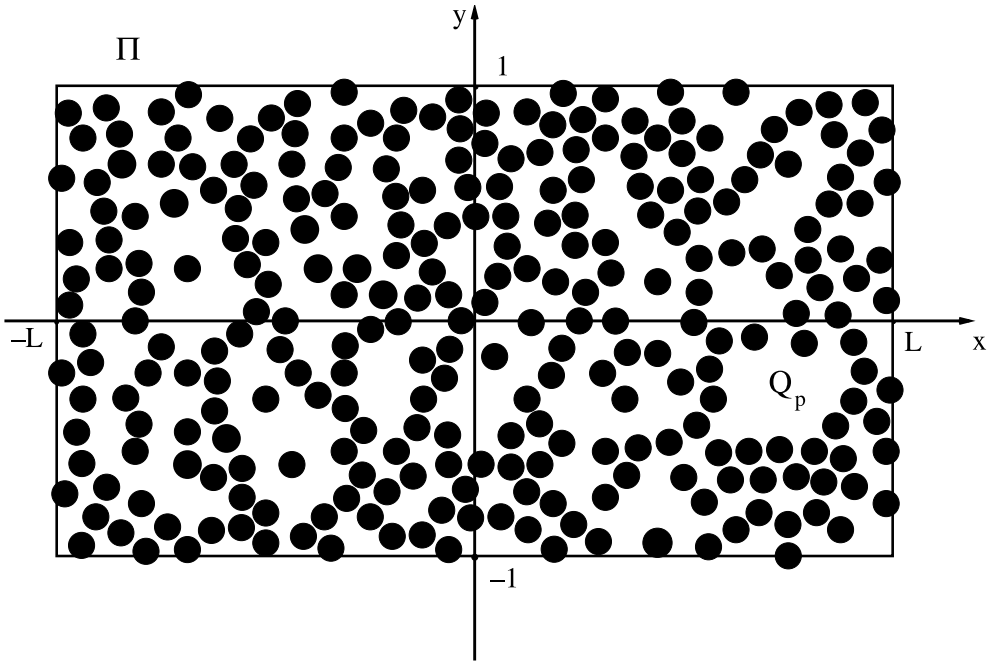


FIG. 2.1. *The composite.*

is the matrix. The potential $\phi(x,y) = \phi(\mathbf{x}), \mathbf{x} = (x,y)$ satisfies

$$\begin{aligned} &\text{(a)} && \Delta\phi = 0 \text{ in } Q_p, \\ &\text{(b)} && \frac{\partial\phi(\pm L, y)}{\partial \mathbf{n}} = 0, \\ &\text{(c)} && \int_{\partial D_i} \partial\phi/\partial\mathbf{n}\, d\mathbf{x} = 0 \text{ for all } i, \\ &\text{(d)} && \phi(\mathbf{x}) = t_i \text{ in } \partial D_i, \\ &\text{(e)} && \phi(x, \pm 1) = \pm 1. \end{aligned} \tag{2.2}$$

We apply the potential ± 1 to the boundaries $y = \pm 1$ (respectively, (2.2)(e)) and assume insulating boundary conditions on the vertical boundaries (2.2)(b). The assumption that the disks are ideally conducting implies (2.2)(c–d), where the constants t_i in (2.2)(d) are arbitrary and they should be determined by solving the system (2.2). The integral condition (2.2)(c) means that the total flux through any disk is zero. If ϕ satisfies (2.2), then the effective conductivity $\hat{a}$ is defined by

$$\hat{a} = \frac{1}{4L} \int_{Q_p} |\nabla\phi|^2 d\mathbf{x}, \tag{2.3}$$

or (see [3])

$$\hat{a} = \frac{1}{2L} \int_{y=1} \nabla\phi \cdot \mathbf{n}\, d\mathbf{x} - \frac{1}{2L} \int_{y=-1} \nabla\phi \cdot \mathbf{n}\, d\mathbf{x} - \frac{1}{4L} \int_{Q_p} |\nabla\phi|^2 d\mathbf{x}, \tag{2.4}$$

because

$$\int_{y=1} \nabla\phi \cdot \mathbf{n}\, d\mathbf{x} - \int_{y=-1} \nabla\phi \cdot \mathbf{n}\, d\mathbf{x} = \int_{Q_p} |\nabla\phi|^2 d\mathbf{x}, \tag{2.5}$$

where $\mathbf{n}$ is the normal to the boundary $y = \pm 1$ and $\int_{y=\pm 1} \nabla\phi \cdot \mathbf{n}\, d\mathbf{x}$ are fluxes through the horizontal boundaries $y = \pm 1$, respectively.

There are two variational definitions (see [3]) of the effective conductivity $\hat{a}$. The first one is given by a minimization problem for the Dirichlet integral

$$\hat{a} = \frac{1}{4L} \min_{\tilde{\phi} \in V_p} \int_{Q_p} |\nabla \tilde{\phi}|^2 d\mathbf{x}, \tag{2.6}$$

where the minimum is taken over a class of all piecewise differentiable potentials $\phi(x, y) \in V_p$, where V_p is defined by

$$V_p = \{\phi \in H^1(Q_p) : \phi(\mathbf{x}) = t_i \text{ on } D_i, \phi(x, \pm 1) = \pm 1\}. \tag{2.7}$$

The Euler–Lagrange equations of the minimization problem (2.6), (2.7) are (2.2). The second variational formulation is given using the dual formulation (2.4):

$$\hat{a} = \frac{1}{2L} \max_{\tilde{\mathbf{v}} \in W_p} \left\{ \int_{y=1} \tilde{\mathbf{v}} \cdot \mathbf{n}\, d\mathbf{x} - \int_{y=-1} \tilde{\mathbf{v}} \cdot \mathbf{n}\, d\mathbf{x} - \frac{1}{2} \int_{Q_p} \tilde{\mathbf{v}}^2 d\mathbf{x} \right\}, \tag{2.8}$$

where the minimum is taken over a class of all fluxes (see [3] for details)

$$W_p = \left\{ \mathbf{v} \in L_2(Q_p) : \mathbf{v}(\pm L, y) \cdot \mathbf{n} = 0, \int_{\partial D_i} \mathbf{v} \cdot \mathbf{n}\, d\mathbf{x} = 0, \nabla \cdot \mathbf{v} = 0 \right\}. \tag{2.9}$$

Hence for any $\phi \in V_p$ and $\mathbf{v} \in W_p$ we have bounds

$$\frac{1}{2L} \left[\int_{y=1} \mathbf{v} \cdot \mathbf{n}\, d\mathbf{x} - \int_{y=-1} \mathbf{v} \cdot \mathbf{n}\, d\mathbf{x} - \frac{1}{2} \int_{Q_p} \mathbf{v}^2 d\mathbf{x} \right] \le \hat{a} \le \frac{1}{4L} \int_{Q_p} |\nabla\phi|^2 d\mathbf{x}. \tag{2.10}$$

Moreover, if $\mathbf{v} = \nabla\phi$, then the upper bound equals the lower bound in (2.10).

2.2. Discrete network. Following [3], we construct the discrete network using the notion of the Voronoi tessellation. We partition the matrix Q_p into simple nonoverlapping geometric figures—necks and triangles. This *triangle-neck partition* is an auxiliary construction, which is used in section 3 as a convenient and efficient tool for the construction of the trial functions for the error estimates.

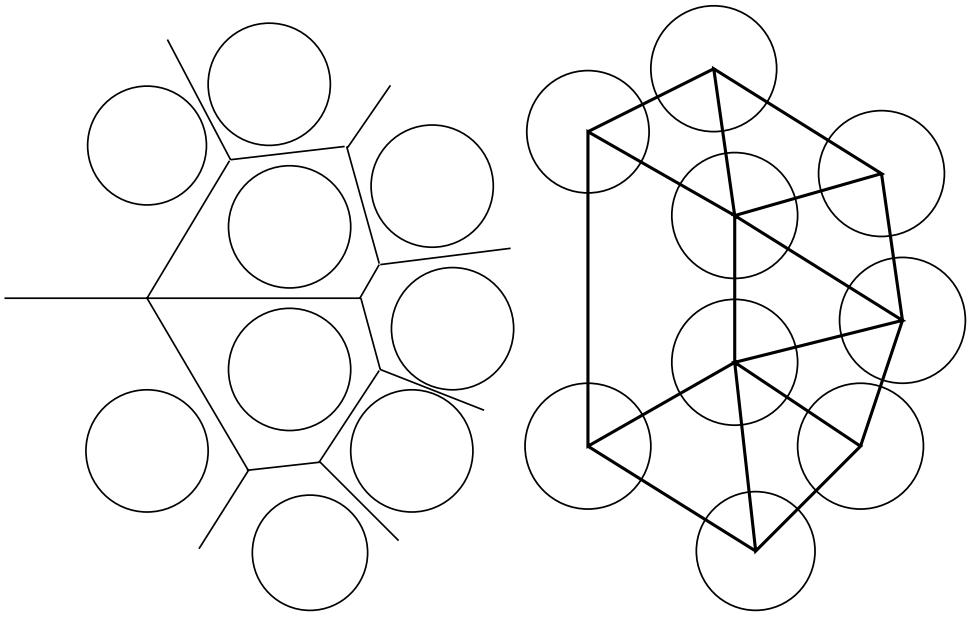
Consider the set of centers x_i, $i = 1, 2, \ldots, N$, of all disks D_i and construct the Voronoi tessellation for the vertices x_i, $i = 1, 2, \ldots, N$.

DEFINITION 2.1. *For a given distribution of vertices x_i, $i = 1, \ldots, N$, in the two-dimensional rectangular domain Π the Voronoi cell V_i associated with x_i is a polygon that consists of all the points in Π at least as close to x_i as to any other vertex. The set of all such Voronoi cells V_i is the Voronoi tessellation of Π.*

Also construct the Delaunay graph (triangulation) dual to the Voronoi tessellation, that is, connect every pair of vertices x_i and x_j by the line segment (edge) e_{ij} if their respective cells share a common edge in the Voronoi tessellation (see Figure 2.2).

DEFINITION 2.2. *Any two disks D_i and D_j are said to be* neighbors *if their centers x_i and x_j are connected by the edge e_{ij}.*

Consider any two neighbors D_i and D_j with the centers x_i and x_j, respectively (see Figure 2.3(a)). Denote by O_n and O_p the endpoints of the common edge of their Voronoi cells V_i and V_j. Then connect the center x_j with all the vertices of its Voronoi



(a) Voronoi tessellation (b) Delaunay graph

FIG. 2.2.

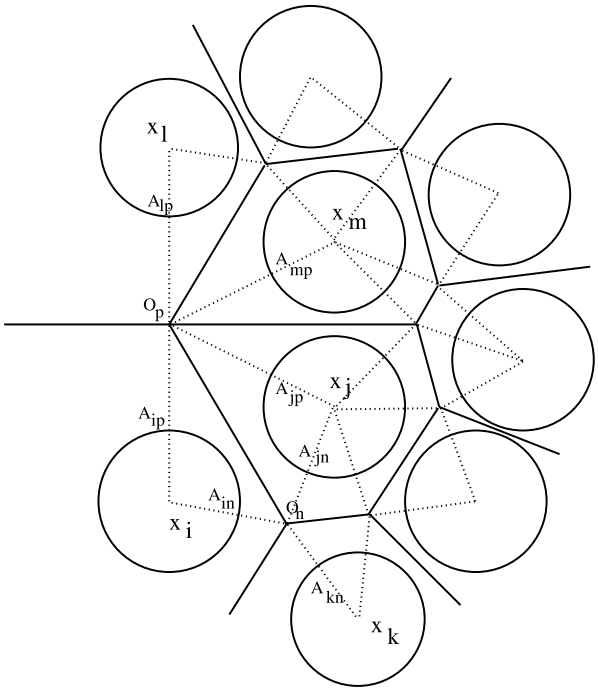
cell V_j by auxiliary line segments (dotted lines in Figure 2.3(a)). Denote by A_{jn} the intersection of the line segment x_jO_n with the circumference of the disk D_j. Finally, define similarly points A_{in} and A_{kn} and connect the points A_{in}, A_{jn}, and A_{kn}.

DEFINITION 2.3. *The* neck Π_{ij} *between the neighbors* D_i *and* D_j *is the curvilinear quadrangle* $A_{in}A_{ip}A_{jp}A_{jn}$*, bounded by the two line segments* $A_{in}A_{jn}$ *and* $A_{ip}A_{jp}$ *and the two arcs* $A_{in}A_{ip}$ *and* $A_{jn}A_{jp}$*.*

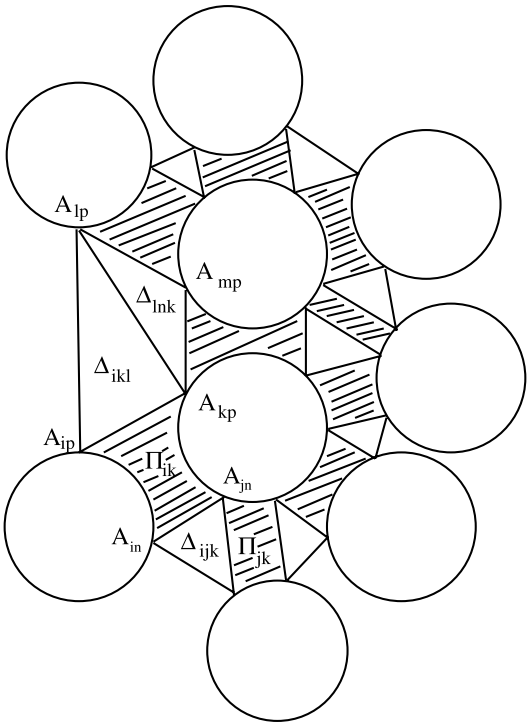
When we apply this algorithm to all neighbors, in general, all these line segments $A_{in}A_{jn}$ partition the domain Q_p into necks Π_{ij} between neighboring disks and triangles Δ_{ijk}. In exceptional cases, instead of triangles we obtain polygons (for example, quadrangle $A_{ip}A_{lp}A_{mp}A_{kp}$ in Figure 2.3(b)) which can be further partitioned into triangles by drawing auxiliary diagonal lines.

The situation at the boundary needs special treatment. For the construction of the partition of Q_p near the boundary $\partial\Pi$ we use reflections about all four parts of $\partial\Pi$. Without loss of generality we assume that all the centers x_i of the disks lie *inside* the domain Π, and hence the centers of the reflected disks will *always* lie outside the domain Π. (The case when the disks lie outside the boundary can also be treated by the model by adding simple but cumbersome modifications.) Consider, for example, the left boundary $x = -L$. The algorithm is shown in Figure 2.4. We reflect symmetrically along the line $x = -L$ *all* the disks, including the disks that intersect the boundary. The latter disks partially overlap with the "ghost" disks (dotted disks in Figure 2.4) which are their mirror images. For the distribution of original disks and the ghost disks we can still apply the Voronoi tessellation and the algorithm proposed for the interior disks. For uniformity of presentation we use, as in [3], a notion of a quasidisk.

DEFINITION 2.4. *A quasidisk* $D_{i''}$*, is the part of a neck* $\Pi_{ii'}$ *that lies on the boundary of* Π*. The radius of a quasidisk is* ∞*.*



(a) Voronoi tessellation



(b) Triangle-neck partition

FIG. 2.3. *Decomposition of a Voronoi cell.*

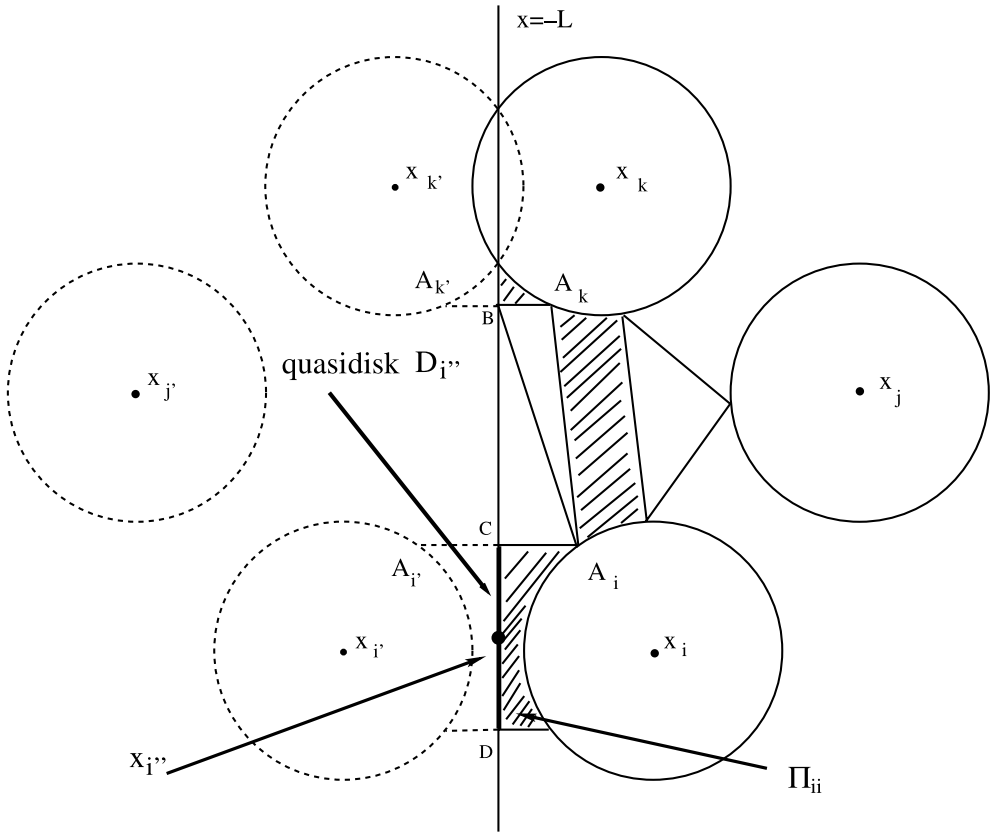


FIG. 2.4. *Vertical boundary.*

An example is the line segment $CD \equiv D_{i''}$ in Figure 2.4. This notion allows us to treat the quasidisks and the original disks uniformly as disks of different radii. In particular, for a quasidisk the definition of a neighbor (Definition 2.2) applies.

DEFINITION 2.5. *The triangle-neck partition $\mathbb{P} = \mathbb{P}(Q_p)$ of the domain Q_p is the set of necks Π_{ij} and triangles Δ_{ijk}.*

The triangle-neck partition is unique up to partitioning of degenerate (exceptional) polygons into triangles. Typically, a neck Π_{ij} is not symmetric with respect to the line connecting the centers of the disks D_i and D_j. An example of a neck is given in Figure 2.5, where we used the local coordinate system when the centers of the both disks lie on the y-axis. In this coordinate system the width of the left half-neck is $|S_1|$, $S_1 < 0$, and the width of the right half-neck is $|S_2|$, $S_2 > 0$. Note that inequalities $S_1 < 0$, $S_2 > 0$ are not true in general, but $S_1 \le S_2$ always by our construction. For uniformity of presentation we view auxiliary diagonals as necks with width zero $S_1 = S_2$. For example, the line segment A_iB in Figure 2.4 corresponds to such a neck.

DEFINITION 2.6. *The maximal and the minimal relative half-neck widths are defined by*

(2.11)
$$\beta_{ij}^{\max} = \max\left(\frac{|S_1|}{R}, \frac{|S_2|}{R}\right), \qquad \beta_{ij}^{\min} = \min\left(\frac{|S_1|}{R}, \frac{|S_2|}{R}\right), 0 \le \beta_{ij}^{\min} \le \beta_{ij}^{\max} < 1.$$

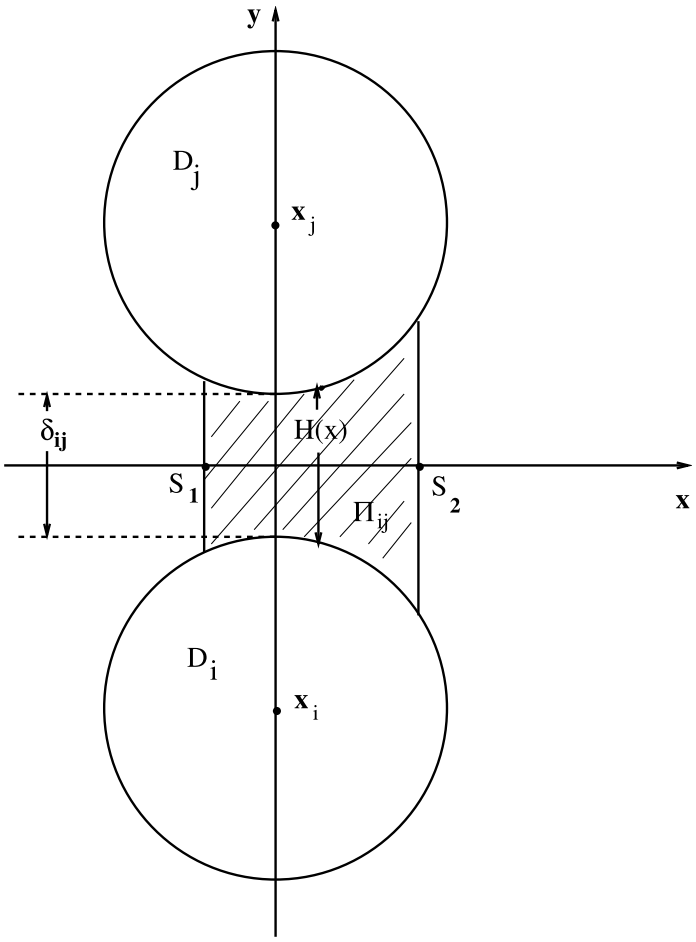


FIG. 2.5. *The hatched region is the neck between two neighbors.*

We use the relative half-neck widths $\beta_{ij}^{\max}$ and $\beta_{ij}^{\min}$ in the error estimates in section 3.

Using the triangle-neck partition we decompose the Dirichlet integral (2.3) into integrals over necks and triangles:

$$\hat{a} = \frac{1}{4L}\int_{Q_p} |\nabla\phi|^2 d\mathbf{x} = \frac{1}{4L}\sum_{\Pi_{ij}} \int_{\Pi_{ij}} |\nabla\phi|^2 d\mathbf{x} + \frac{1}{4L}\sum_{\Delta_{ijk}} \int_{\Delta_{ijk}} |\nabla\phi|^2 d\mathbf{x}. \tag{2.12}$$

The DNA of the effective conductivity is based on the observation in [13] that for high concentration of the disks the fluxes $\nabla\phi$ are significant only in necks Π_{ij} between closely spaced disks,

$$\sum_{\Delta_{ijk}} \int_{\Delta_{ijk}} |\nabla\phi|^2 d\mathbf{x} \ll \sum_{\Pi_{ij}} \int_{\Pi_{ij}} |\nabla\phi|^2 d\mathbf{x}, \tag{2.13}$$

and in these necks the fluxes can be easily computed (as in [13]) by the linear interpolation between the values of the potentials on the disks. More specifically, if we

align the neck Π_{ij} between two neighbors D_i and D_j with the vertical direction as indicated in Figure 2.5, then by the linear interpolation the local flux in Π_{ij} satisfies

$$\nabla\phi = \left(0, \frac{t_i - t_j}{H(x)}\right), \tag{2.14}$$

$$H(x) = \begin{cases} \delta_{ij} + 2R - 2\sqrt{R^2 - x^2} & \text{for disks,} \\ \delta_{ij} + R - \sqrt{R^2 - x^2} & \text{for quasidisks,} \end{cases}$$

where δ_{ij} is given by the following definition.

DEFINITION 2.7. *The length δ_{ij} of a neck Π_{ij} is*

$$\delta_{ij} = \begin{cases} d_{ij} - 2R & \text{if } D_i \text{ and } D_j \text{ are disks,} \\ d_{ij} - R & \text{if one of } D_i \text{ is a quasidisk,} \end{cases} \tag{2.15}$$

where $d_{ij} = |x_i - x_j|$ is the (Euclidean) distance between x_i and x_j.

Using (2.14)

$$\int_{\Pi_{ij}} |\nabla\phi|^2 d\mathbf{x} = \int_{S_1}^{S_2} \frac{(t_i - t_j)^2}{H^2(x)} H(x)dx = g_{ij}(t_i - t_j)^2, \tag{2.16}$$

where

$$g_{ij} = \int_{S_1}^{S_2} \frac{dx}{H(x)}, \tag{2.17}$$

and t_i, t_j are potentials on D_i, D_j, respectively, and S_1, S_2 are as defined in Figure 2.5. If we use (2.16) with (2.17) for all neighbors, then observation (2.13) would imply

$$\hat{a} \equiv \frac{1}{4L} \min_{\phi \in V_p} \int_{Q_p} |\nabla\phi|^2 d\mathbf{x} \sim I \equiv \frac{1}{4L} \min_{\mathbf{t}} \sum_{\Pi_{ij}} g_{ij}(t_i - t_j)^2, \tag{2.18}$$

where $\mathbf{t} = (t_1, t_2, \ldots, t_N)$.

This is our *modified network approximation*. The *energy of the discrete network*

$$I = \frac{1}{4L} \min_{\mathbf{t}} \sum_{\Pi_{ij}} g_{ij}(t_i - t_j)^2 \tag{2.19}$$

is determined by the choice of the *specific fluxes* g_{ij}. Following [13] they were chosen in [3] to be

$$g_{ij} = \pi \sqrt{\frac{2R_iR_j}{R_i + R_j}} \Bigg/ \sqrt{\delta_{ij}} \tag{2.20}$$

for *closely spaced* disks and zero otherwise. Both formulas (2.20) and (2.17) give the same leading term in the power expansion in $\delta_{ij} \to 0$ (see [3], [4] for details), and hence these formulas are asymptotically equivalent as $\delta_{ij} \to 0$. We propose here to use (2.17), instead of (2.20), and use it for *all* neighbors not necessarily closely spaced. Such *modified* choice of the specific fluxes allows us to derive tight variational bounds for the relative error of our modified network approximation. The choice (2.17) validates the approximation (2.18) for much more general nonuniform irregular distributions, when a significant fraction of the inclusions does not participate in the

conducting spanning cluster, whereas in [3] it was shown that (2.20) validates the approximation (2.18) only for arrays of inclusions in which all neighbors are closely spaced ("randomized hexagonal arrays"). The main goal of this paper is to give sufficient conditions for these nonuniform irregular distributions, under which (2.18) is a valid approximation, and provide a rigorous a priori estimate on the relative error of this approximation. The full details of our construction of the modified network approximation can be found in [4].

The relation (2.18) is a DNA, because we approximate the continuum minimization problem with a discrete minimization problem of a quadratic form (2.19). The unknowns of the minimization problem are the values of the discrete potentials t_i on the interior disks, and the quadratic form is defined on a graph (network) where the vertices x_i are the centers of the disks D_i and the edges e_{ij} are the necks Π_{ij}. This graph is the Delaunay triangulation for the centers of the disks, modified by an additional construction at the boundaries. The construction is as follows. If for a vertex x_i its Voronoi cell V_i is adjacent to the boundary (that is, one of the sides of V_i lies on the boundary of Π) and the radius of the disk $R_i = R$ is smaller than the distance from x_i to this boundary, then connect x_i with this boundary by the perpendicular line. Denote the intersection of this perpendicular and the boundary by $x_{i''}$, $i'' > N$, and the line segment between x_i and $x_{i''}$ by the edge $e_{ii''}$. This modification added vertices $x_{i''}$, $i'' > N$, that lie on the boundary of Π (see Figure 2.4). These vertices are in one-to-one correspondence with quasidisks in the triangle-neck partition.

The discrete potentials $\tilde{t}_i$ at the "boundary" vertices x_i, are prescribed on the horizontal boundaries by

$$\tilde{t}_i = \pm 1 \text{ for } x_i \in S^{\pm}, \tag{2.21}$$

where $S^{\pm}$ are the upper/lower boundary vertices defined as follows.

DEFINITION 2.8. *A vertex x_i is an upper (lower/left/right) boundary vertex, if x_i lies on the upper (lower/left/right) boundary (an added vertex $x_{i''}$, $i'' > N$) or x_i is the center of the disk D_i, that intersects the upper (lower/left/right) boundary. The set S^+ ($S^-/S^l/S^r$) is the set of upper (lower/left/right) boundary vertices.*

The minimization problem (2.19)–(2.21) amounts to solving the corresponding linear algebraic system that determines the discrete potentials $\tilde{t}_i$ at the "interior" vertices $x_i \in \mathbb{I}$, $\mathbb{I} = \{x_i, i = 1, \dots, N\} \setminus (S^+ \cup S^-)$. A discrete version of insulating boundary conditions on $S^l \cup S^r$ can be formulated as follows. If a vertex of the discrete network $x_{i''} \notin \mathbb{I} \cup (S^+ \cup S^-)$, then $x_{i''}$ must be a center of a *quasidisk* that lies on the left or the right boundary $x_{i''} \in S^l \cup S^r$. For such vertices $\tilde{t}_{i''} = \tilde{t}_i$, where $\tilde{t}_i$ and $\tilde{t}_{i''}$ are the values of the potential on the disks $D_i \in \mathbb{I} \cup (S^+ \cup S^-)$ and $D_{i''}$, respectively, where D_i is the uniquely determined disk connected with $D_{i''}$ (e.g., in Figure 2.4 $D_{i''} \equiv CD$). Therefore,

$$\sum_{\Pi_{ik}, i'' \text{ fixed}} g_{i''k}(t_{i''} - t_k) = g_{i''i}(t_{i''} - t_i) = 0 \quad \text{for all } x_{i''} \notin \mathbb{I} \cup (S^+ \cup S^-). \tag{2.22}$$

DEFINITION 2.9. *For a given distribution of disks D_i with centers x_i, $i = 1, \dots, N$, the discrete network $\mathbb{D}$ is a set of vertices x_i, $i = 1, \dots, M$, $M \geq N$, and edges e_{ij} between neighbors x_i and x_j of the modified Delaunay graph.*

By [3] a necessary condition for the validity of (2.18) is the existence of a conducting spanning cluster.

DEFINITION 2.10. *For any discrete network $\mathbb{D}$ (or any subgraph $\mathbb{D}'$ of it) a spanning cluster is the (unique) connected component of $\mathbb{D}$ (or $\mathbb{D}'$) that contains at least one path that connects the S^+ and S^- and at least one path that connects S^l and S^r.*

The spanning cluster is conducting if the distance δ_{ij} between every two consecutive vertices of this graph $\delta_{ij} \leq \delta$, where δ is sufficiently small. A sufficient condition for the existence of the conducting spanning cluster is the δ-**N** connectedness property of the discrete network $\mathbb{D}$ which can be formulated in terms of δ-subgraphs of $\mathbb{D}$.

DEFINITION 2.11. *For any $\delta > 0$ the δ-subgraph $\mathbb{D}_\delta$ of the discrete network (graph) $\mathbb{D}$ is the subset of edges e_{ij} and their incident vertices x_i and x_j of $\mathbb{D}$ such that their length $\delta_{ij} \leq \delta$. For any subgraph a vertex is incident if it is an end-vertex of one of its edges.*

The δ-**N** connectedness property of $\mathbb{D}$ is used extensively in this paper; therefore, for completeness, in the rest of this section we give the precise graph-theoretical definitions related to this notion. Most of them are taken from [5].

DEFINITION 2.12. *A path of a graph $\mathbb{D}$ from x_0 to x_n is an alternating sequence of*

$$x_0, e_{01}, x_1, e_{12}, x_2, e_{23}, \ldots, x_{n-1}, e_{n-1n}, x_n$$

of distinct vertices x_i and edges e_{ij}. Such a path has size $\mathbf{n}$*, and the vertices x_0 and x_n are said to be the end-vertices. The vertices $x_1, \ldots, x_{n-1}$ are said to be the interior vertices.*

Conventionally (see [5]), the *size* of a path is called its *length*. Here we do not use this standard notation, because the following definition of the length of a path is more natural in our setting (due to Definition 2.7).

DEFINITION 2.13. *The length of a path is the sum of lengths of its edges e_{ij}. The length of an edge e_{ij} is the length of the corresponding neck Π_{ij} as in Definition* 2.7.

DEFINITION 2.14. *An internal cycle C of $\mathbb{D}$ is an alternating sequence of*

$$x_0, e_{01}, x_1, e_{12}, x_2, e_{23}, \ldots, x_n, e_{n0}, x_0$$

of vertices and edges, such that

$$x_0, e_{01}, x_1, e_{12}, x_2, e_{23}, \ldots, x_n$$

is a path, and e_{n0} connects the vertices x_0 and x_n. Such a cycle has size $\mathbf{n}+1$.

DEFINITION 2.15. *A boundary cycle C of $\mathbb{D}$ is a path*

$$x_0, e_{01}, x_1, e_{12}, x_2, e_{23}, \ldots, x_n$$

such that the end-vertices x_0 and x_n lie on the boundary of the domain Π. Such a boundary cycle has size $\mathbf{n}$.

Note that the end-vertices x_0 and x_n of a boundary cycle may belong to different boundaries, for example S^- and S^l, respectively.

DEFINITION 2.16. *A minimal cycle $C_{\min}$ is an (internal or boundary) cycle such that for any two vertices x_i and x_j of this cycle the shortest path from x_i to x_j is a subset of the cycle $C_{\min}$. If the cycle $C_{\min}$ is a boundary cycle, we also require that for any interior point x_i of this cycle the shortest path from to x_i to any point x_k on the boundary is a subset of the cycle $C_{\min}$.*

Note that we require for a minimal boundary cycle an additional condition. This condition guarantees that a boundary cycle cannot be shortened by connecting an

interior point of this cycle with the boundary. Definition 2.16 is a formalization of an intuitive notion of a hole in a composite. Each hole in a composite is surrounded by a loop of conducting disks (see Figure 1.1(a)). On the modified Delaunay graph this loop corresponds to an **N**-gon, which is a minimal cycle of this graph.

DEFINITION 2.17. *The size* **N** *of the largest minimal cycle of a (sub)graph* $\mathbb{D}_\delta$ *is the upper bound on the size of all its minimal cycles; that is,*

$$\mathbf{N} = \max_{C_{\min} \subset \mathbb{D}_\delta} size(C_{\min}), \tag{2.23}$$

where $C_{\min}$ *is a minimal cycle.*

The interior of a cycle C, denoted Int_C, is a (closed) polygon having the cycle C as its boundary; that is, $\partial \mathrm{Int}_C = C$. However, the definition of the interior of a boundary cycle requires an additional technical construction. Naturally, the interior of a boundary cycle can be defined as a (closed) polygon having the cycle C and some parts of the boundary of the domain Π as its boundary. However, there are exactly two such polygons such that their union is the whole domain Π. Among these two polygons we choose for the interior the one with the smallest area.

The *degree* of the connectedness of the whole graph $\mathbb{D}$ can now be quantified in terms of the two parameters, δ and **N**, and an a priori relative error estimate for the DNA $\mathbb{D}$ is determined in terms of these parameters only.

DEFINITION 2.18. *For a fixed* δ *a discrete network (graph)* $\mathbb{D}$ *is* δ-**N** *connected if*

(i) *the* δ*-subgraph* $\mathbb{D}_\delta$ *contains the spanning cluster as in Definition* 2.10,

(ii) *the size of the largest minimal cycle of* $\mathbb{D}_\delta$ *is* **N**.

In this paper we are interested in δ-**N** connected discrete networks such that the size of the composite is large, compared to the perimeter of the largest minimal cycle:

$$(2R + \delta)\mathbf{N} < \min(2, 2L). \tag{2.24}$$

If (2.24) holds, then (i) is equivalent to the following:

(i') For every point y of the domain Π there exists a minimal cycle $C_{\min}$ of the δ-subgraph $\mathbb{D}_\delta$ such that $y \in \mathrm{Int}_{C_{\min}}$; and if this minimal cycle is a boundary cycle, then its end-vertices either lie on the same (left/right/upper/lower) part of the boundary $\partial\Pi$ or they lie on two adjacent parts of $\partial\Pi$, e.g., left and upper.

The condition (i') is technical, but in the proofs of our main results we use (i') instead of (i). Thus, let us show that these conditions are equivalent. To argue by contradiction we assume that (2.24) holds, but there are no paths on the graph $\mathbb{D}_\delta$ that connect S^+ and S^-. Then there exists a path in the *whole* domain Π that connects S^l and S^r and does not intersect $\mathbb{D}_\delta$. The length of this path, on the one hand, must be larger than the distance between the left and the right boundaries; on the other hand, it cannot be larger than the diameter of the largest minimal cycle. Recall that the distance between any two vertices in $\mathbb{D}$ does not exceed $2R + \delta$, and hence the inequality $2L \le (2R + \delta)\mathbf{N}$ must hold which contradicts (2.24). Suppose now (i) holds. Then the spanning cluster partitions Π into polygons ("holes in the cluster"). Therefore, every point $y \in \Pi$ lies in one of them. These polygons can be chosen to be nonoverlapping and so that their boundaries are interior minimal cycles of $\mathbb{D}_\delta$, or boundary minimal cycles of $\mathbb{D}_\delta$ and some parts of $\partial\Pi$. Hence it is left to check that every boundary minimal cycle satisfies the condition that its end-vertices either lie on the same part of the boundary $\partial\Pi$ or they lie on two adjacent parts of $\partial\Pi$. This follows from (2.24).

Finally, we note that the existence of a path connecting S^+ and S^- implies the existence of a path connecting S^l and S^r. Indeed, suppose (2.24) holds, but there is

no path in $\mathbb{D}_\delta$ that connects S^l and S^r. Then the connected component of $\mathbb{D}_\delta$ that contains the path from S^+ to S^- is not connected to one of the boundaries S^l or S^r. Then there exists a *boundary minimal cycle* in $\mathbb{D}_\delta$ that connects S^+ and S^-. On the one hand, its length is at most $(2R+\delta)\mathbf{N}$; on the other hand, it must exceed 2, the distance between the upper and the lower boundaries. This is impossible, because it contradicts (2.24).

2.3. Properties of the discrete network. Here we collect some results on the properties of the discrete network. The first lemma gives an upper bound on the number of necks and triangles that lie in the interior of any minimal cycle of the δ-subgraph in terms of the size of largest minimal cycle $\mathbf{N}$. Consider a δ-$\mathbf{N}$ connected discrete network $\mathbb{D}$ (Definition 2.18). Let us now consider a minimal cycle $C_{\min}$ of the δ-subgraph $\mathbb{D}_\delta$. Denote by $\#\Delta_{C_{\min}}$ the number of the triangles Δ_{ijk} that lie in the interior of this minimal cycle $\Delta_{ijk} \subset \mathrm{Int}_{C_{\min}}$. Similarly, $\#\Pi_{C_{\min}}$ is the number of the necks Π_{ij} that lie in the interior of this minimal cycle $\Pi_{ij} \subset \mathrm{Int}_{C_{\min}}$, and $\#x_{C_{\min}}$ is the number of the vertices (centers of disks) x_i such that $x_i \subset \mathrm{Int}_{C_{\min}}$.

LEMMA 2.19. *Suppose the discrete network $\mathbb{D}$ is δ-$\mathbf{N}$ connected. Then for any minimal cycle $C_{\min}$ of the δ-subgraph $\mathbb{D}_\delta$ the number of triangles $\#\Delta_{C_{\min}}$, and the number of necks $\#\Pi_{C_{\min}}$ that lie in the interior of $C_{\min}$, satisfy the bounds*

$$\#\Delta_{C_{\min}} \le 2\left(\mathbf{N} + \frac{2}{\pi\sqrt{3}}\mathbf{N}^2\right), \quad \#\Pi_{C_{\min}} \le 3\left(\mathbf{N} + \frac{2}{\pi\sqrt{3}}\mathbf{N}^2\right). \tag{2.25}$$

The proof is basically the isoperimetric inequality together with the Euler's formula. For details see [4].

LEMMA 2.20. *There is a unique solution $\mathbf{t} = \{t_i | x_i \in \mathbb{I}\}$ of the discrete minimization problem* (2.19)–(2.21). *This solution satisfies a discrete analogue of Euler–Lagrange equations (compare to* (2.2)*)*

$$\sum_{\Pi_{ik}, i \text{ fixed}} g_{ik}(t_i - t_k) = 0 \quad \text{for all } x_i \in \mathbb{I}. \tag{2.26}$$

Proof. A solution that satisfies (2.26) exists, because the quadratic form (2.19) is positive definite. The discrete network is a connected graph in the sense that there is a path between each vertex x_i and a boundary vertex $x_j \in S^\pm$. This implies that the solution is unique. □

Similar to the fluxes through the horizontal boundaries on the right-hand side of (2.4), denote by P^+ and P^- the discrete fluxes through the boundaries S^+ and S^-, respectively;

$$P^+ \equiv \sum_{\Pi_{ij}, x_i \in S^+} g_{ij}(t_i - t_j), \qquad P^- \equiv \sum_{\Pi_{ij}, x_i \in S^-} g_{ij}(t_i - t_j). \tag{2.27}$$

Then

$$\frac{1}{4L}(P^+ - P^-) = \frac{1}{4L}\sum_{\Pi_{ij}} g_{ij}(t_i - t_j)^2 \equiv I(\mathbf{t}). \tag{2.28}$$

Formula (2.28) is a discrete analogue of (2.5). For the proof see [3].

LEMMA 2.21 (discrete maximum principle). *Suppose $\mathbf{t} = \{t_1, t_2, \dots, t_M\}$ is the solution of the $\mathbb{D}$ problem* (2.19)–(2.21). *For any (internal or boundary) cycle C of $\mathbb{D}$ define*

$$t_{\max} = \max(t_i), x_i \in C, \qquad t_{\min} = \min(t_i), x_i \in C.$$

Then for any vertex x_k with potential t_k such that $x_k \in \mathrm{Int}_C$, that is, x_k belongs to the interior of the cycle C (as in Definition 2.18), we have

$$t_{\min} \le t_k \le t_{\max}.$$

The proof of the discrete maximum principle is by contradiction; it is fairly standard and it could be found in [3]. As a corollary of the discrete maximum principle we have the following lemma.

LEMMA 2.22. *If the discrete network $\mathbb{D}$ is δ-$\mathbf{N}$ connected, then for any minimal cycle $C_{\min}$ of the δ-subgraph $\mathbb{D}_\delta$ and a vertices $x_k \in \mathrm{Int}_{C_{\min}}$ and $x_l \in \mathrm{Int}_{C_{\min}}$ we have*

$$(t_k - t_l)^2 \le \mathbf{N} \sum_{\Pi_{ij} \in C_{\min}} (t_i - t_j)^2. \tag{2.29}$$

Proof. By the discrete maximum principle

$$(t_k - t_l)^2 \le (t_{\max} - t_{\min})^2, \tag{2.30}$$

where

$$t_{\max} = \max(t_i), x_i \in C_{\min}, \qquad t_{\min} = \min(t_i), x_i \in C_{\min}.$$

Suppose the maximum $t_{\max}$ and the minimum $t_{\min}$ are achieved at the vertices $x' \in C_{\min}$ and $x'' \in C_{\min}$, respectively. Since both vertices belong to the minimal cycle $C_{\min}$, there is a part of this minimal cycle which is a path with the size $\le \mathbf{N}$ that connects them. Therefore, by the triangle inequality for the values of the potentials t_i, $x_i \in C$

$$(t_{\max} - t_{\min})^2 \le \mathbf{N} \sum_{\Pi_{ij} \in C_{\min}} (t_i - t_j)^2$$

which inserted in (2.30) yields (2.22). □

LEMMA 2.23. *Suppose*

$$|g^0_{ij} - g_{ij}| \le C_0 \text{ as } \delta_{ij} \to 0, \tag{2.31}$$

and $I^0(\mathbf{t^0})$ is another DNA with the specific fluxes g^0_{ij} and the energy I^0. Then the bound $\frac{|\hat{a}-I|}{I} \le C_1\sqrt{\delta/R}$ implies $\frac{|\hat{a}-I^0|}{I^0} \le C_2\sqrt{\delta/R}$, where C_2 depends on C_0 and C_1 only.

Lemma 2.23 gives an equivalence of the energy $I(\mathbf{t})$ to the energy of any other discrete network that uses a set of specific fluxes g^0_{ij} that satisfies (2.31). Since (see derivation in [4]) both formulas (2.20) and (2.17) give the same leading term in the power expansion in $\delta_{ij} \to 0$, the previous lemma implies that for the purpose of estimating the effective conductivity of a composite where inclusions are almost touching our model and the model introduced in [3] are equivalently good as $\delta \to 0$. For the proof see [4], which, in particular, shows that under the δ-$\mathbf{N}$ packing condition the relative error of the discrete network I^0 introduced in [3] (when (2.17) are replaced by (2.20)) satisfies

$$\frac{|\hat{a} - I^0|}{I^0} \le 9.82\mathbf{N}^4\sqrt{\frac{\delta}{R}}.$$

3. Variational error estimates.

3.1. The lower bound. Following [3] the test function $\mathbf{v}$ for the lower bound is chosen to be zero everywhere except the necks Π_{ij} between adjacent disks; however, in our case for the specific fluxes we use (2.17) instead of (2.20).

PROPOSITION 3.1. *The lower bound on $\hat{a}$ in terms of g_{ij} and the parameters of the solution of the discrete minimization problem is*

$$I(\mathbf{t}) = \frac{1}{4L} \sum_{\Pi_{ij}} g_{ij}(t_i - t_j)^2 \le \hat{a}, \tag{3.1}$$

where $\mathbf{t} = \{t_1, t_2, \dots, t_M\}$ are the values of the discrete potentials of the solution of the discrete minimization problem (2.19)–(2.21).

The left-hand side in (3.1) is always positive, which reflects the physics of the problem. The analogous lower bound in Proposition 2.1 in [3] is positive and sufficiently tight for the δ-3 close packing condition only. Our bound allows us to handle general distribution of disks that satisfy the δ-$\mathbf{N}$ close packing condition for any $\mathbf{N}$.

Proof. Consider two neighbors centered at x_i and x_j. Suppose the potentials t_i and t_j on the disks D_i and D_j are the solutions to the minimization problem (2.19)–(2.21). Rotate the domain so that in the local coordinates the neck between them is aligned with the direction of the y-axis (as in Figure 2.5). Define

$$\mathbf{v} = \begin{cases} (0, \frac{t_i - t_j}{H(x)}) & \text{in the neck } \Pi_{ij}, \\ (0,0) & \text{otherwise}, \end{cases} \tag{3.2}$$

where $H(x)$ is the distance between the disks. Since for a piecewise constant function the divergence-free condition amounts to checking that the normal components of $\mathbf{v}$ match along the discontinuity, we see that our trial function $\mathbf{v}$ is divergence-free. The matching condition $\int_{\partial D_i} \mathbf{v} \cdot \mathbf{n} d\mathbf{x} = 0$ is satisfied (as in [3]) due to (2.17) and (2.26). The insulating condition $\mathbf{v}(\pm L, y) \cdot \mathbf{n} = 0$ at the vertical boundary is satisfied by (2.22). Hence $\mathbf{v} \in W_p$. Observe that for the trial function (3.2) the fluxes through the upper and the lower boundary of Π are exactly the discrete fluxes P^+ and P^-:

$$P^+ = \int_{y=1} \mathbf{v} \cdot \mathbf{n}\, d\mathbf{x}, \qquad P^- = \int_{y=-1} \mathbf{v} \cdot \mathbf{n}\, d\mathbf{x}.$$

Following [3] we have

$$\int_{Q_p} \mathbf{v}^2 d\mathbf{x} = \sum_{\Pi_{ij}} g_{ij}(t_i - t_j)^2,$$

because $\mathbf{v} \equiv 0$ on every triangle and the Dirichlet integral over a neck

$$\int_{\Pi_{ij}} \mathbf{v}^2 d\mathbf{x} = (t_i - t_j)^2 \int_{S_1}^{S_2} \frac{dx}{H(x)} = g_{ij}(t_i - t_j)^2.$$

Using (2.28) we have

$$\frac{1}{2L} \left[\int_{y=1} \mathbf{v} \cdot \mathbf{n}\, d\mathbf{x} - \int_{y=-1} \mathbf{v} \cdot \mathbf{n}\, d\mathbf{x} - \frac{1}{2} \int_{Q_p} \mathbf{v}^2 d\mathbf{x} \right] = I(\mathbf{t}).$$

By the first inequality in (2.10) we have (3.1). ∎

3.2. The upper bound.

PROPOSITION 3.2. *The upper bound on $\hat{a}$ in terms of g_{ij} and the parameters of the solution of the discrete minimization problem is*

$$\hat{a} \le \frac{1}{4L}\sum_{\Pi_{ij}}[g_{ij}+C_{ij}](t_i-t_j)^2 = I(\mathbf{t}) + \frac{1}{4L}\sum_{\Pi_{ij}} C_{ij}(t_i-t_j)^2, \tag{3.3}$$

where all $C_{ij} = C_{ij}(\beta_{ij}^{\max})$ depend only on the relative neck thicknesses $\beta_{ij}^{\max}$ defined (2.11). *Moreover, if*

$$\beta_{ij}^{\max} \le \beta < 1, \tag{3.4}$$

then $C_{ij} \le C$ with some $C = C(\beta)$.

It was shown in [4] that

$$C_{ij} \le \frac{|\ln(1-\beta_{ij}^{\max})| + \pi + \ln 2}{6} + \frac{4}{\sqrt{1-[\beta_{ij}^{\max}]^2}}.$$

Proof. Consider a piecewise continuous test function ϕ. Similar to [3], the function ϕ is linear in y in the neck Π_{ij} with the values t_i and t_j on the boundary of the disks ∂D_i and ∂D_j (Figure 2.5). Then on the neck Π_{ij}

$$\phi(x,y) = t_i + \frac{(t_j-t_i)(y+H(x)/2)}{H(x)} = t_i + (t_j-t_i)\left[\frac{y}{H(x)}+\frac{1}{2}\right]$$
$$\text{for } y \in \left[-\frac{H(x)}{2}, \frac{H(x)}{2}\right]. \tag{3.5}$$

The function ϕ is linear in the Δ_{ijk} (see Figure 3.1) with the values t_i^0, t_j^0, and t_k^0 at the vertices of Δ_{ijk}. In Figure 3.1 these vertices are the points A, B, and C, respectively. In a neck Π_{ij}

$$\int_{\Pi_{ij}} (\partial\phi/\partial y)^2 dxdy = \int_{S_1}^{S_2} \frac{(t_j-t_i)^2}{H^2(x)} H(x)dx = (t_j-t_i)^2 g_{ij} \tag{3.6}$$

and

$$\int_{\Pi_{ij}} (\partial\phi/\partial x)^2 dxdy = \int_{\Pi_{ij}} (t_j^0-t_i^0)^2 \left[\frac{yH'(x)}{H^2(x)}\right]^2 dxdy$$

$$= \frac{1}{12}(t_j^0-t_i^0)^2 \int_{S_1}^{S_2} \frac{(H')^2(x)}{H(x)}dx.$$

Since $H(x) = \delta + 2R - 2\sqrt{R^2-x^2}$, therefore $H'(x) = \frac{2x}{\sqrt{R^2-x^2}}$, and

$$\frac{H'^2(x)}{H(x)} = \frac{2R}{R^2-x^2} + \frac{2}{\sqrt{R^2-x^2}}.$$

Hence as in [3]

$$\int_{\Pi_{ij}} (\partial\phi/\partial x)^2 dxdy \le C(\beta_{ij}^{\max})(t_j-t_i)^2 \le C(t_j-t_i)^2 \tag{3.7}$$

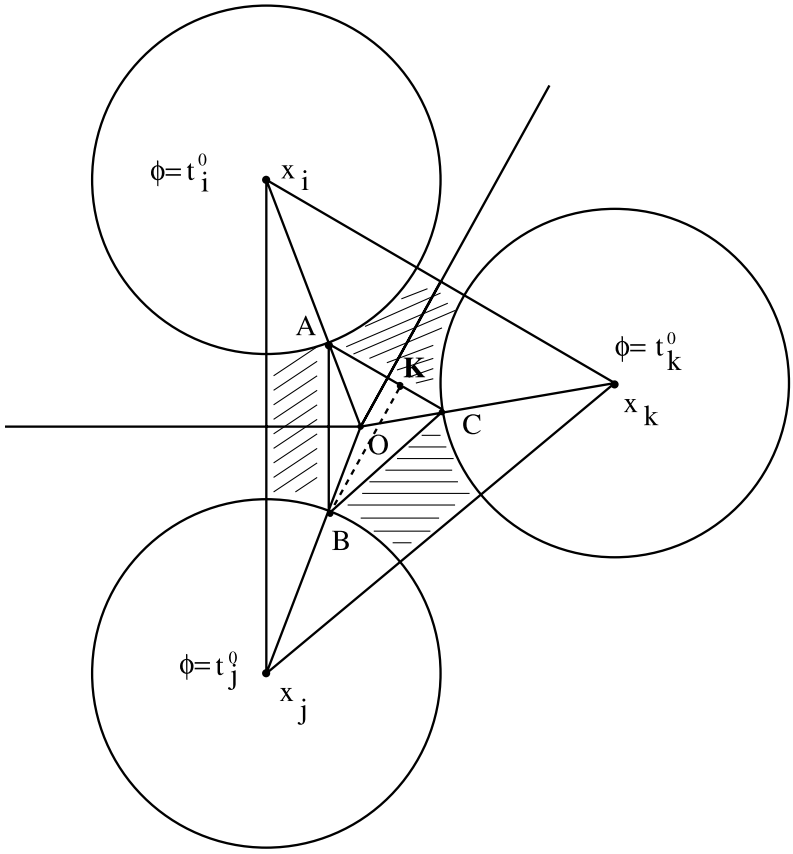


FIG. 3.1. *Typical $\Delta ABC \equiv \Delta_{ijk}$; half-necks are hatched.*

if $\beta_{ij} \le \beta < 1$. Each triangle is bounded by three necks. In Figure 3.1 we show a typical case with three disks centered at x_i, x_j, and x_k, three half-necks, and the $\Delta ABC \equiv \Delta_{ijk}$ bounded by these half-necks. Suppose in ΔABC the side $|AC|$ is the longest, and the side $|BC|$ is the shortest. Then (see details in [4])

$$\int_{ABC} |\nabla\phi|^2 d\mathbf{x} \le \frac{2}{\sin(\angle BAC)}\Big((t_i - t_j)^2 + (t_k - t_i)^2\Big).$$

Since

$$\frac{1}{\sin(\angle BAC)} = \frac{1}{\sin(\pi/2 - \angle x_k x_j O)} \le \frac{1}{\sqrt{1 - (\beta_{jk}^{\max})^2}},$$

therefore

$$\int_{ABC} |\nabla\phi|^2 d\mathbf{x} \le C(\beta_{jk}^{\max})((t_i - t_j)^2 + (t_k - t_i)^2) \le C((t_i - t_j)^2 + (t_k - t_i)^2) \tag{3.8}$$

if $\beta_{jk}^{\max} \le \beta < 1$. Combining (3.6), (3.7), (3.8) and summing over all necks and triangles we have (3.3). □

3.3. The error estimate.

THEOREM 3.3. *If for a distribution of disks D_i, $i = 1, \dots, N$, its discrete network is δ-$\mathbf{N}$ connected, then the relative error*

$$\frac{|\hat{a} - I|}{I} \le C(\mathbf{N})\sqrt{\frac{\delta}{R}}, \tag{3.9}$$

where $\hat{a}$, I are defined in (2.3), (2.19)*, respectively, and*

$$C(\mathbf{N}) \le 2.56\mathbf{N}^4. \tag{3.10}$$

Inequality (3.10) is a very rough upper bound, and it could be significantly improved by more careful analysis. In particular, if $\mathbf{N} = 3$, then $C(\mathbf{3}) \le 3.84$; if $\mathbf{N} = 4$, then $C(\mathbf{4}) \le 12.73$. For simplicity of presentation we do not specify our numerical estimate (3.10). For more details see [4].

Proof. As $\delta_{ij} \to 0$, we know that $g_{ij} = O(\delta^{-1/2})$ and therefore if there is a conducting spanning cluster, then (see [3]) $I(\mathbf{t}) = O(\delta^{-1/2})$. Hence (3.10) will follow if we can "absorb" *all* the $O(1)$ terms in (3.3) as the "smaller order corrections" into $O(\sqrt{\frac{R}{\delta}})$ terms and derive an estimate

$$I(\mathbf{t}) \le \hat{a} \le I(\mathbf{t}) + O(1) \le I(\mathbf{t})\left(1 + C\sqrt{\frac{\delta}{R}}\right), \tag{3.11}$$

where the first inequality in (3.11) follows from (3.1). The second inequality is also immediate if *all* necks are short: $\delta_{ij} \le \delta$. If not all the necks are short, then $g_{ij} = O(1)$ and therefore C_{ij} in (3.3) are compatible in magnitude with g_{ij}. Here we use the assumption that our discrete network is δ-$\mathbf{N}$ connected. The key observation here is that the centers of any two neighbors D_k and D_l must lie inside a minimal cycle $C_{\min}$. Therefore, by Lemma 2.22 the values of the potentials t_k and t_l at these disks satisfy

$$(t_k - t_l)^2 \le \mathbf{N} \sum_{\Pi_{ij} \in C_{\min}} (t_i - t_j)^2.$$

By Lemma 2.19 we have a bound, that depends on $\mathbf{N}$ only, on the total number of vertices, necks, and triangles that lie inside any minimal cycle $C_{\min}$. Therefore, for our trial function ϕ for the upper bound, the Dirichlet integral over the all the triangles and necks that lie inside any $C_{\min}$ is a smaller order correction to the energy of the minimal cycle $C_{\min}$:

$$\sum_{\Pi_{ij}, e_{ij} \in C_{\min}} g_{ij}(t_i - t_j)^2.$$

Thus we have

$$\hat{a} \le \frac{1}{4L}\sum_{\Pi_{ij}} [g_{ij} + C](t_i - t_j)^2 \le \frac{1}{4L} \sum_{\Pi_{ij}, e_{ij} \notin \mathbb{D}_\delta} g_{ij}(t_i - t_j)^2$$

$$+ \frac{1}{4L} \sum_{\Pi_{ij}, e_{ij} \in \mathbb{D}_\delta} [g_{ij} + C(\mathbf{N})](t_i - t_j)^2 \le \frac{1}{4L} \sum_{\Pi_{ij}, e_{ij} \notin \mathbb{D}_\delta} g_{ij}(t_i - t_j)^2$$

$$+ \frac{1}{4L} \sum_{\Pi_{ij}, e_{ij} \in \mathbb{D}_\delta} g_{ij} \left(1 + C(\mathbf{N})\sqrt{\frac{\delta}{R}}\right)(t_i - t_j)^2 \le \left(1 + C(\mathbf{N})\sqrt{\frac{\delta}{R}}\right) \frac{1}{4L} \sum_{\Pi_{ij}} g_{ij}(t_i - t_j)^2.$$

Hence we have proved (3.11). ∎

3.4. A posteriori numerical error. The main goal of this paper is to give a rigorous quantitative justification of discrete network model (2.18) by means of a priori error estimates. However, the use of our trial functions for the upper and the lower bounds also provides us with a numerical a posteriori error. We must solve the discrete network problem (2.19)–(2.21), construct the trial functions $\phi \in V_p$ (see section 3.1) and $\mathbf{v} \in W_p$ (see section 3.2), and evaluate explicitly the left-hand side and the right-hand side of the upper and lower bound (2.10). The evaluation of this dual bound is not computationally expensive, because we use simple trial functions—they are given by explicit analytic formulas in the necks (for derivation see [4]):

$$g_{ij} = \int_{S_1}^{S_2} \frac{dx}{H(x)}$$

$$= \left[-\alpha + 2\frac{\delta_{ij} + R}{\sqrt{\delta_{ij}^2 + 2R\delta_{ij}}} \arctan\left(\frac{\sqrt{\delta_{ij}^2 + 2R\delta_{ij}}\tan(\alpha/2)}{\delta_{ij}} \right) \right] \Bigg|_{\arcsin(S_1/R)}^{\arcsin(S_2/R)},$$

and they are linear interpolations on the triangles. Also the use of the a posteriori error widens the range of the characteristic distance δ, where the discrete network gives a good approximation. This section implements this idea for numerical simulations of a randomized hexagonal lattice. For details see [4].

Our numerical experiments consist of three parts: numerical simulations of a randomized hexagonal distribution of disks; numerical evaluations of the dual variational bounds; and the statistics.

The distribution of disks is implemented by randomization of a periodic hexagonal lattice of disks of equal radii $R_i = R = 0.02$ on a square domain $[-1, 1] \times [-1, 1]$, with a volume fraction f, and then removal of some fraction f_r of these disks from this distribution. For fixed f and f_r this algorithm creates a distribution of disks with the volume fraction $f_0 = f - f_r$.

For a given distribution of disks we compute $I = I(\mathbf{t})$ (formula (2.19)), the energy of the discrete network. After the energy I is computed, we also compute the trial function ϕ for the upper bound as in section 3.2, and then we compute $I_\phi = \frac{1}{4L}\int_{Q_p} |\nabla\phi|^2 d\mathbf{x}$ for this trial function. Therefore, by construction in sections 3.1 and 3.2 we have $I \le \hat{a} \le I_\phi$. Hence I and I_ϕ are a posteriori lower and upper bounds, respectively, for the effective conductivity of a composite with a given distribution of disks.

The simulations are done with the 0.05 increments of f_r. For fixed f and f_r there were 80 simulations. For the mean we use the notation $\mathbb{E}(I) = 1/n \sum_{k=1}^{n} I^k$, where I^k is the result of kth simulation with fixed f and f_r and the number of simulations $n = 80$. Here we present the results of the numerical simulations that show the dependence of the effective conductivity on the presence of holes in the matrix.

In Figure 3.2 we plot $\mathbb{E}(I)$ (solid lines) and $\mathbb{E}(I_\phi)$ (dotted lines) as functions of the volume fraction $f_0 = 0.105, \ldots, 0.905$. For the lower two lines the volume fraction of removed disks $f_r = 0$ is fixed, for the upper two lines the initial volume fraction $f = 0.905$ is fixed, and $f_r = f - f_0$. If the total volume fraction of the inclusions is fixed, then the increase of the volume fraction of holes in the material implies that the interparticle distance δ decreases. Hence the percolation effects play a more significant role. Indeed, we observe that for the *same* volume fraction $f_0 = 0.605$ a distribution of disks with holes has the effective conductivity at least two times larger

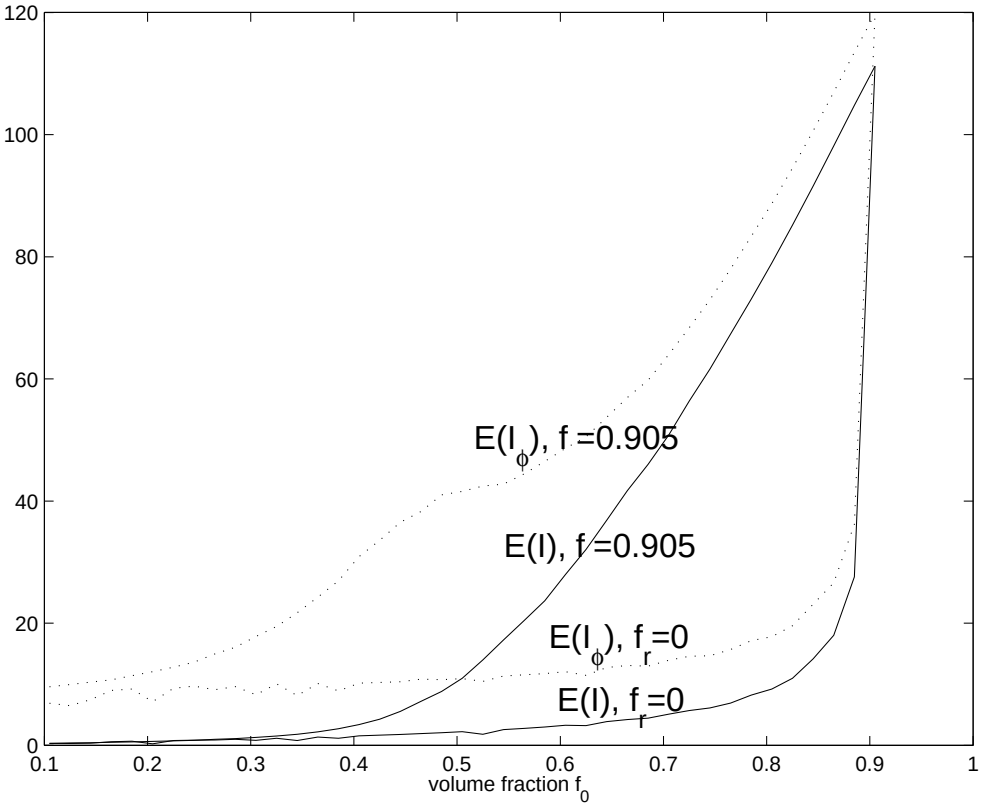


FIG. 3.2. $\mathbb{E}(I)$ *(solid lines) and* $\mathbb{E}(I_\phi)$ *(dotted lines) as functions of the volume fraction* $f_0 = 0.105, \ldots, 0.905$. *Lower two lines for the case with no holes. Upper two lines for the case with holes.*

than a distribution of disks without holes. (The lower bound in the first case is more than two times larger than the upper bound in the composite with no holes.)

When $f_0 \leq .35$ the effective conductivities of a material with holes and of a material without holes are numerically very close, at least the computed a posteriori error does not allow us to distinguish between these composites. For such volume fractions there are no percolation effects in both cases.

Observe that in the presence of holes the a posteriori error of the network approximation $I_\phi - I$ is significantly larger than in the case when there are no holes. For example, when $f_0 = .5$ in the presence of holes the a posteriori error is 3.5 times larger than the error in the case when there are no holes; however, for the same volume fraction of inclusions the *relative* a posteriori error estimate is better in the presence of holes. For example, when $f = .6$ the relative a posteriori error estimate in the presence of holes in the conducting cluster is up to 8 times better than for uniformly highly packed composites. Hence, we observe numerically that our network approximation works better for irregular geometric patterns, which are not quasi-hexagonal, that is, when a typical number of nearest neighbors can vary significantly.

Finally, let us compare the a posteriori error estimates discussed in this section and the a priori error estimates given in section 3.3 for a quasi-hexagonal case, $\mathbf{N} = 3$. Since $C(3) \leq 3.84$ in (3.9), a straightforward calculation shows that 10% accuracy is achieved when δ is about 1500 and 250 times smaller than the disk's radius R for the

a priori and a posteriori estimates, respectively.

4. Conclusions. In this paper we introduced and justified the modified network approximation which generalizes the network approximation from [3]. This modified network approximation is no longer asymptotic in nature. We decompose the Dirichlet integral (2.6) for the effective conductivity into two parts: the network approximation, which is a quadratic form, and the error term. Our modified network approximation accounts for *all* fluxes between the neighboring particles, where neighbors are defined via Voronoi tessellation. If the fluxes are small (neighbors are not closely spaced), then the corresponding coefficients in the modified network approximation g_{ij} are not significant, but unlike [3] we do not need to introduce any cut-off distance in the numerical implementation. For the network approximation the error term is explicitly estimated based on the construction of the trial functions using the triangle-neck partition of the domain. The approach developed in this paper allowed us to consider generic geometrical arrays of the particles which satisfy the so-called δ-$\mathbf{N}$ close packing condition. This condition allows for strongly nonuniform geometrical arrays when a significant fraction of the particles does not participate in the conducting cluster. We have shown numerically how such nonuniformity affects the error estimate and therefore the quality of the approximation. We observe that irregularity in the geometrical distribution of the inclusions in all simulations consistently lead to a significant (up to 10 times) increase in the effective conductivity at the same total volume fraction of the inclusions.

Thus we conclude that our approximation provides a very efficient computational tool for evaluation of the effective properties of high contrast composites, which is capable of capturing the effects of irregular geometrical arrays with a good control of the approximation error. We expect that the method developed in this work will be generalized for more complicated problems of highly packed elastic and fluid composites.

Acknowledgment. We are grateful to Ivo Babuska for raising the question of the error estimate.

REFERENCES

[1] V. Ambegaokar, B. I. Halperin, and J. Slanger, *Hopping conductivity in disordered systems*, Phys. Rev. B, 4 (1971), pp. 2612–2620.

[2] N. S. Bakhvalov and G. P. Panasenko, *Homogenization: Averaging Processes in Periodic Media*, Kluwer, Dordrecht, The Netherlands, 1989.

[3] L. Berlyand and A. Kolpakov, *Network approximation in the limit of small interparticle distance of the effective properties of a high contrast random dispersed composite*, Arch. Ration. Mech. Anal., 159 (2001), pp. 179–227.

[4] L. Berlyand and A. Novikov, *Error of the Network Approximation for Densely Packed Composites with Irregular Geometry*, IMA preprint series 1849, 2002.

[5] B. Bollobás, *Modern Graph Theory*, Grad. Texts in Math. 184, Springer-Verlag, New York, 1998.

[6] L. Borcea, *Asymptotic analysis of quasi-static transport in high contrast conductive media*, SIAM J. Appl. Math., 59 (1998), pp. 597–635.

[7] L. Borcea, J. G. Berryman, and G. Papanicolaou, *Matching pursuit for imaging high contrast conductivity*, Inverse Problems, 15 (1999), pp. 811–849.

[8] L. Borcea and G. C. Papanicolaou, *Network approximation for transport properties of high contrast materials*, SIAM J. Appl. Math., 58 (1998), pp. 501–539.

[9] L. Borcea and G. Papanicolaou, *Low frequency electromagnetic fields in high contrast media*, in Surveys on Solution Methods for Inverse Problems, D. Colton, H. W. Engl, A. Louis, J. R. McLaughlin, and W. Rundell, eds., Springer, Vienna, New York, 2000.

[10] O. Bruno, *The effective conductivity of strongly heterogeneous composites*, Proc. Roy. Soc. London Ser. A, 433 (1991), pp. 353–381.
[11] J. P. Clerc, G. Giraud, J. M. Laugier, and J. M. Luck, *The electrical conductivity of binary disordered systems, percolation clusters, fractals and related models*, Adv. in Phys., 39 (1990), pp. 191–309.
[12] B. I. Halperin, *Remarks on percolation and transport in networks with a wide range of bond strengths*, Phys. D, 38 (1989), pp. 179–183.
[13] J. B. Keller, *Conductivity of a medium containing a dense array of perfectly conducting spheres or cylinders or nonconducting cylinders*, J. Appl. Phys., 34 (1963), pp. 991–993.
[14] J. Koplik, *Creeping flow in two-dimensional networks*, J. Fluid Mech., 119 (1982), pp. 219–247.
[15] S. M. Kozlov, *Geometrical aspects of averaging*, Russian Math. Surveys, 44 (1989), pp. 91–144.
[16] R. McPhedran, *Transport property of cylinder pairs of the square array of cylinders*, Proc. Roy. Soc. London Ser. A, 408 (1986), pp. 31–43.
[17] R. McPhedran, L. Poladian, and G. W. Milton, *Asymptotic studies of closely spaced, highly conducting cylinders*, Proc. Roy. Soc. London Ser. A, 415 (1988), pp. 185–196.
[18] L. M. Schwartz, J. R. Banavar, and B. I. Halperin, *Biased-diffusion calculations of effective transport in inhomogeneous continuum systems*, Phys. Rev. B (3), 40 (1989), pp. 9155–9161.
[19] C. B. Shah and Y. C. Yortsos, *The permeability of strongly-disordered systems*, Phys. Fluids, 8 (1996), pp. 280–282.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 409–434

PHASE-FIELD MODELS WITH HYSTERESIS IN ONE-DIMENSIONAL THERMOVISCOPLASTICITY*

PAVEL KREJČÍ†, JÜRGEN SPREKELS‡, AND ULISSE STEFANELLI§

Abstract. This paper introduces a combined one-dimensional model for thermoviscoplastic behavior under solid-solid phase transformations that incorporates the occurrence of hysteresis effects in both the strain-stress law and the phase transition described by the evolution of a *phase-field* (which is usually closely related to an order parameter of the phase transition). Hysteresis is accounted for using the mathematical theory of *hysteresis operators* developed in the past thirty years. The model extends recent works of the first two authors on phase-field models with hysteresis to the case when mechanical effects can no longer be ignored or even prevail. It leads to a strongly nonlinear coupled system of partial differential equations in which hysteresis nonlinearities occur at several places, even under time and space derivatives. We show the thermodynamic consistency of the model, and we prove its well-posedness.

Key words. phase-field systems, phase transitions, hysteresis operators, thermoviscoplasticity, thermodynamic consistency

AMS subject classifications. 34C55, 35K60, 47J40, 74K05, 74N30, 80A22

PII. S0036141001387604

1. Introduction and physical motivation.

In this paper, we study initial-boundary value problems for systems of partial differential equations of the form

$$\rho u_{tt} - \mu u_{xxt} = \sigma_x + f(x,t) \quad \text{a.e. in } \Omega_T, \tag{1.1}$$

$$\sigma = \mathcal{H}_1[u_x, w] + \theta\, \mathcal{H}_2[u_x, w] \quad \text{a.e. in } \Omega_T, \tag{1.2}$$

$$\Big(C_V\theta + \mathcal{F}_1[u_x, w]\Big)_t - \kappa\theta_{xx} = \mu u_{xt}^2 + \sigma u_{xt} + g(x,t,\theta) \quad \text{a.e. in } \Omega_T, \tag{1.3}$$

$$\nu w_t = -\psi \quad \text{a.e. in } \Omega_T, \tag{1.4}$$

$$\psi = \mathcal{H}_3[u_x, w] + \theta\, \mathcal{H}_4[u_x, w] \quad \text{a.e. in } \Omega_T, \tag{1.5}$$

$$u(\cdot,0) = u_0, \quad u_t(\cdot,0) = u_1, \quad \theta(\cdot,0) = \theta_0, \quad w(\cdot,0) = w_0 \quad \text{a.e. in } \Omega, \tag{1.6}$$

$$u(0,t) = 0, \quad \mu\, u_{xt}(1,t) + \sigma(1,t) = 0, \quad \theta_x(0,t) = \theta_x(1,t) = 0 \quad \text{a.e. in } (0,T), \tag{1.7}$$

*Received by the editors April 9, 2001; accepted for publication (in revised form) April 5, 2002; published electronically December 3, 2002.
http://www.siam.org/journals/sima/34-2/38760.html
†Mathematical Institute, Academy of Sciences of the Czech Republic, Žitná 25, CZ–11567 Praha 1, Czech Republic (krejci@math.cas.cz).
‡Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, D-10117 Berlin, Germany (sprekels@wias-berlin.de).
§Università degli Studi di Pavia, Dipartimento di Matematica, Via Ferrata 1, I–27100 Pavia, Italy (ulisse@dimat.unipv.it).

where $\Omega := (0,1)$, $T > 0$ denotes some final time, and where $\Omega_t := \Omega \times (0,t)$ for $t \in (0,T]$.

The system (1.1)–(1.7) constitutes a model for the one-dimensional thermomechanical developments in a linearly viscous piece of wire of unit length in which a solid-solid phase transition takes place. In this connection, the unknowns u, θ, σ, w, ψ denote displacement, absolute (Kelvin) temperature, elastoplastic stress, phase variable (usually a so-called *generalized freezing index*; cf. [15]), and the thermodynamic force driving the phase transition, respectively. The positive physical constants ρ, μ, C_V, κ, ν denote mass density, viscosity, specific heat, heat conductivity, and a relaxation coefficient, in that order. For the sake of notational convenience, we will always assume without loss of generality that $\rho = \mu = C_V = \kappa = \nu = 1$. Finally, the expressions $\mathcal{H}_j$, $1 \le j \le 4$, and $\mathcal{F}_1$ are nonlinearities of *hysteresis type* (to be specified below).

The equations (1.1), (1.3), (1.4) represent the equation of motion, the balance of internal energy, and the phase evolution equation, in that order (see below); equation (1.2) is the constitutive law relating strain and phase variable to the elastoplastic stress, and (1.4) expresses that the phase variable evolves into the opposite direction of the thermodynamic force driving the phase transition. In addition, the boundary conditions (1.7) indicate that the wire is thermally insulated at both ends, fixed at $x = 0$, and stress-free at $x = 1$.

The motivation to study systems of the above type is twofold. On the one hand, it is well known that for many materials the macroscopic strain-stress (ε-σ, where $\varepsilon = u_x$ is the linearized strain and u is the displacement) relations measured in uniaxial load-deformation experiments strongly depend on the absolute (Kelvin) temperature θ and, at the same time, exhibit a strong elastoplasticity witnessed by the occurrence of *hysteresis loops* that are *rate-independent*, i.e., independent of the speed with which there are traversed. Due to the hysteresis, which reflects the presence of a *rate-independent memory* in the material, the stress-strain relation can no longer be expressed in terms of a simple single-valued function. Among the materials showing very strong temperature-dependent hysteretic effects are the so-called *shape memory alloys* (see Figure 1 below and Chapter 5 in [2]); but even quite ordinary steels are well known to exhibit this kind of behavior (cf. [21]), although to a smaller extent.

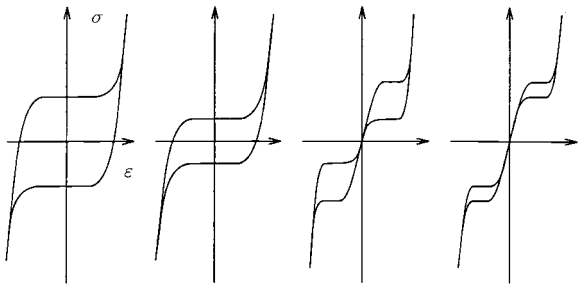


FIG. 1. *Schematic load-deformation curves in shape memory alloys, with temperature increasing from left to right.*

Usually the occurrence of a hysteresis in the macroscopic stress-strain relations is accompanied (or even triggered) by changes between different configurations of the crystal lattice within the solid. It thus makes sense to complement macroscopic equations of thermoelastoplasticity by field equations accounting for such phase trans-

formations on the micro- and/or mesoscales.

On the other hand, phase transition phenomena are often accompanied by macroscopic hysteresis effects that are caused by thermal and/or mechanical stresses acting on the micro- and/or mesoscales. It then makes sense to complement the field equations describing the macroscopic phase transition by equations modeling such micro- or mesoscopic stresses.

A classical approach to such problems would be the following. One first tries to construct a local *free energy function* of the form

$$F(\varepsilon, w, \theta) = \theta(1 - \log(\theta)) + F_1(\varepsilon, w) + \theta\, F_2(\varepsilon, w) \tag{1.8}$$

in such a way that the experimentally observed ε-σ and/or w-ψ hysteresis loops are approximately matched using the relations

$$\sigma = \frac{\partial F}{\partial \varepsilon}(\varepsilon, w, \theta), \quad \psi = \frac{\partial F}{\partial w}(\varepsilon, w, \theta), \tag{1.9}$$

then determines the corresponding *internal energy* U and *entropy* S,

$$\begin{aligned} U(\varepsilon, w, \theta) &:= F(\varepsilon, w, \theta) - \theta\, \frac{\partial F}{\partial \theta}(\varepsilon, w, \theta) = \theta + F_1(\varepsilon, w), \\ S(\varepsilon, w, \theta) &:= -\frac{\partial F}{\partial \theta}(\varepsilon, w, \theta) = \log(\theta) - F_2(\varepsilon, w), \end{aligned} \tag{1.10}$$

and finally inserts these expressions in the governing field equations: equation of motion,

$$u_{tt} - \tilde{\sigma}_x = f \quad (\tilde{\sigma} = \text{total stress} = \sigma + \text{viscous stress}), \tag{1.11}$$

balance of internal energy,

$$U_t - \theta_{xx} = \tilde{\sigma} u_{xt} + g \quad (U = \text{internal energy}), \tag{1.12}$$

and phase evolution equation (1.4).

We then obtain (1.1), (1.3), (1.4) if we put

$$\begin{aligned} &\mathcal{H}_1[\varepsilon, w] := \frac{\partial F_1}{\partial \varepsilon}(\varepsilon, w), \quad \mathcal{H}_2[\varepsilon, w] := \frac{\partial F_2}{\partial \varepsilon}(\varepsilon, w), \\ &\mathcal{H}_3[\varepsilon, w] := \frac{\partial F_1}{\partial w}(\varepsilon, w), \quad \mathcal{H}_4[\varepsilon, w] := \frac{\partial F_2}{\partial w}(\varepsilon, w), \\ &\mathcal{F}_1[\varepsilon, w] := F_1(\varepsilon, w). \end{aligned} \tag{1.13}$$

In order that an ε-σ (or w-ψ, respectively) hysteresis be modeled by (1.9), $F(\cdot, w, \theta)$ ($F(\varepsilon, \cdot, \theta)$, respectively) needs to be a nonconvex function within the range of interesting temperatures.

This approach has advantages: if the nonlinearities involved in (1.1)–(1.7) are smooth functions, then the vast literature on one-dimensional thermoviscoelasticity (we just refer to the fundamental papers [4], [5]) can be applied to derive results concerning well-posedness and asymptotic behavior. However, while this approach is capable of correctly predicting many of the experimentally observed phenomena, it also has certain disadvantages from the phenomenological (engineering) point of view: the use of a nonconvex free energy does not guarantee that a hysteresis actually occurs; one only observes that unstable branches are traversed with a very high speed, which

may look like a hysteresis if the speed is interpreted as infinite. For the moment, there is no mathematical theory which would allow us to rigorously justify this singular transition and give a correct account of the inherent memory structures that are responsible for the complicated loopings in the interior of the external hysteresis loops that are observed in experiments.

To avoid these difficulties, the first two authors have recently proposed a different approach using the theory of *hysteresis operators* developed in the past thirty years (let us at least refer to the monographs [8], [20], [22], [2], [9] devoted to this subject). In this approach, we replace the relations (1.9) by the identities (1.2), (1.5), where the expressions $\mathcal{H}_j$, $1 \le j \le 4$, and $\mathcal{F}_1$ are no longer real-valued *functions* but *hysteresis operators* acting between suitable function spaces. This approach has been successfully carried out for the two cases when either we have one-dimensional thermoelastoplastic hysteresis without the order parameter evolution (that is, we have (1.1)–(1.3) with no dependence on w; cf. the papers [10], [11]) or we have a multidimensional phase transition without mechanical effects (that is, we have (1.3)–(1.5) with no dependence on u, σ; see [7], [12], [13], [14], [15], [16], [17]). In this paper, we want to extend some of these results to the fully coupled problem.

At this point we remark that in [10], [11] a more general constitutive law than (1.2) has been treated in that there the hysteresis operators could also depend on θ. However, no dependence on w was admitted in [10], [11]. It is in fact this additional dependence that forces us to assume σ in the form (1.2) which, on the other hand, is quite typical in the Landau theory of phase transitions (cf. Chapter 4 in [2]). The extension of the results of this paper to more general temperature-dependent hysteresis operators seems to be a very difficult unsolved problem.

We also note that in [18] the present authors have studied a related version of system (1.1)–(1.7): there, an additional curvature term was added on the left of (1.1), and the boundary conditions for u were of the form $u(0,t) = u(1,t) = u_{xx}(0,t) = u_{xx}(1,t) = 0$. Let us stress the fact that the mathematical analysis carried out in [18] differs considerably from that used in this paper. In fact, the problem investigated here is more difficult than that studied in [18]: if the smoothing term γu_{xxxx} is present in (1.1), then already the first a priori estimate (see below) yields an $L^\infty(L^2)$-bound for the strain gradient u_{xx} and thus an $L^\infty(L^\infty)$-bound for u_x. This means that the Andrews transformation (see (3.1) below) used in this paper to eventually bound u_x is not needed. As a consequence, the solution established in this paper enjoys less smoothness than that in [18]. On the other hand, our analysis does not apply to the case when (1.1) is complemented with zero boundary conditions at both ends of the wire: we have to assume a stress-free regime at one of the ends in order to take advantage of the Andrews transformation.

Let us now recall some basic facts about the notion of hysteresis operators (for details, we refer to the monographs mentioned above). Let $T > 0$ denote some (final) time. A mapping $\mathcal{H}$ from the set $\mathrm{Map}[0,T] := \{w : [0,T] \to \mathbb{R}\}$ into itself is called a *hysteresis operator* if it is *causal*, that is, if for all $w_1, w_2 \in \mathrm{Map}[0,T]$ and $t \in [0,T]$ we have the implication

$$w_1(\tau) = w_2(\tau) \quad \forall\, \tau \in [0,t] \;\Rightarrow\; \mathcal{H}[w_1](t) = \mathcal{H}[w_2](t),$$

and if it is *rate-independent*, that is, if for every $w \in \mathrm{Map}[0,T]$ and every continuous increasing mapping α of $[0,T]$ onto $[0,T]$ we have

$$\mathcal{H}[w \circ \alpha](t) = \mathcal{H}[w](\alpha(t)) \quad \forall\, t \in [0,T].$$

In the case of partial differential equations, when the input functions not only depend on a time variable $t \in [0,T]$ but also on a space variable $x \in [0,1]$, it is necessary to extend the above notion. In this situation, it is natural to associate with a hysteresis operator $\mathcal{H}$ defined on $\mathrm{Map}[0,T]$ in the above sense an operator $\hat{\mathcal{H}}$ acting on $\mathrm{Map}([0,1] \times [0,T])$ by simply putting

$$\hat{\mathcal{H}}[w](x,t) := \mathcal{H}[w(x,\cdot)](t). \tag{1.14}$$

It is customary to identify the operators $\mathcal{H}$ and $\hat{\mathcal{H}}$. The hysteresis operators appearing in (1.1)–(1.7) have to be understood in this way.

The advantage of this approach is that an operator equation like (1.2), (1.5), is suited much better than a relation like (1.9) to keep track of the memory effects imprinted on the material in the past history. In fact, the output at any time $t \in [0,T]$ may depend on the whole evolution of the input in the time interval $[0,t]$. Observe that the rate-independence implies that the hysteresis behavior cannot be expressed in terms of an integral operator of convolution type; i.e., we are not dealing with a model with fading memory.

Unfortunately, there are also disadvantages: the input-output behavior of hysteresis operators usually cannot be described explicitly, and they have, as a rule, only very restricted smoothness properties. In fact, nontrivial hysteresis operators are, as a rule, *not differentiable*, but at best only (possibly locally) *Lipschitz continuous* in suitable function spaces; in addition, they carry a *nonlocal memory* with respect to time.

Both nondifferentiability and presence of a memory are unpleasant features from the mathematical point of view. For instance, the classical method of deriving higher order a priori estimates for w (namely, differentiation of (1.4) with respect to t and testing with w_t) does not immediately work, since there is no chain rule for the hysteretic nonlinearities; also, we may not simply differentiate (1.2) or (1.5) with respect to x. These facts result in a lack of compactness and thus in difficulties in existence proofs.

However, hysteresis operators usually dissipate energy which typically is proportional to the area of closed traversed loops in the hysteresis diagram. Let us explain this fact for one fundamental hysteresis operator which plays a most prominent role in the theory, namely the so-called *stop operator* or *Prandtl's normalized elastic–perfectly plastic element*. To this end, let $r > 0$ (the yield limit) and $\sigma_r^0 \in [-r,r]$ (the initial stress) be given. For any input function $\varepsilon \in W^{1,1}(0,T)$, we define the output $\sigma_r \in W^{1,1}(0,T)$ as the solution to the variational inequality (the index t denotes time differentiation)

$$\sigma_r(t) \in [-r,r] \quad \forall\, t \in [0,T], \quad \sigma_r(0) = \sigma_r^0, \tag{1.15}$$

$$(\varepsilon_t(t) - \sigma_{r,t}(t))(\sigma_r(t) - \eta) \geq 0 \quad \forall\, \eta \in [-r,r] \quad \text{a.e. in } (0,T). \tag{1.16}$$

In Figure 2, the typical input-output behavior is depicted.

It can easily be proved (see, for instance, [9], where also the multidimensional case is treated) that (1.15)–(1.16) admits a unique solution $\sigma_r \in W^{1,1}(0,T)$ for every $\varepsilon \in W^{1,1}(0,T)$ and $\sigma_r^0 \in [-r,r]$. The corresponding solution operator

$$s_r : [-r,r] \times W^{1,1}(0,T) \to W^{1,1}(0,T) : (\sigma_r^0, \varepsilon) \mapsto \sigma_r \tag{1.17}$$

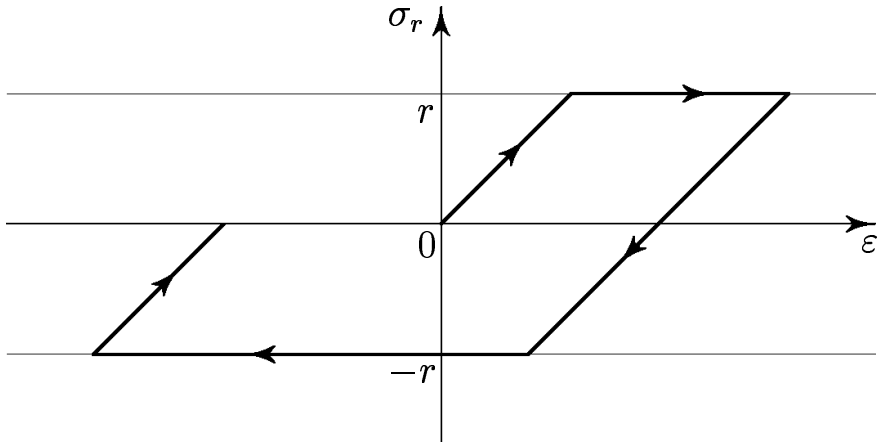


FIG. 2. *Prandtl's normalized elastic–perfectly plastic element.*

is just the stop operator. It has the well-known property (cf. [2], [9]) that for any $r_1, r_2 \in [0, +\infty)$, $\sigma^0_{r_j} \in [-r_j, +r_j]$, $j = 1, 2$, $t \in [0, T]$, and $\varepsilon_1, \varepsilon_2 \in C[0, T]$, it holds that

$$\begin{aligned} &\big|s_{r_1}[\sigma^0_{r_1}, \varepsilon_1](t) - s_{r_2}[\sigma^0_{r_2}, \varepsilon_2](t)\big| \\ &\le |\varepsilon_1(t) - \varepsilon_2(t)| + \max\Big\{|r_1 - r_2| + \max_{0\le\tau\le t} |\varepsilon_1(\tau) - \varepsilon_2(\tau)|, \big|\sigma^0_{r_1} - \sigma^0_{r_2}\big|\Big\}. \end{aligned} \tag{1.18}$$

In addition, it holds for any $\varepsilon \in W^{1,1}(0, T)$ that

$$\big\|s_r[\sigma^0_r, \varepsilon]\big\|_\infty \le r, \quad s_r[\sigma^0_r, \varepsilon]_t(t) = \varepsilon_t(t) \quad \text{if} \quad \big|s_r[\sigma^0_r, \varepsilon](t)\big| < r, \tag{1.19}$$

$$\big|s_r[\sigma^0_r, \varepsilon]_t\big|^2 = s_r[\sigma^0_r, \varepsilon]_t \varepsilon_t \quad \text{a.e. in } (0, T). \tag{1.20}$$

For each fixed initial condition σ^0_r, the stop operator is a hysteresis operator in the sense of the above definition, where the value of σ^0_r accounts for some previous memory. On the other hand, the value at $t = 0$ of the output of any rate-independent operator can be represented as a function of the initial value of the input. For the operators occurring in our model, we will not make any special assumptions about their initial configurations, except for those that follow from the general analytic conditions imposed below in hypothesis (H4), like, e.g., Lipschitz continuous input-output relation.

We now describe the intrinsic dissipation property of the stop operator. It results if we insert $\eta = 0$ in (1.16). We then obtain that the energy $\mathcal{P}_r := \frac{1}{2} s_r^2$ of the stop element satisfies the inequality

$$\frac{d}{dt}\, \mathcal{P}_r[\sigma^0_r, \varepsilon](t) \le s_r[\sigma^0_r, \varepsilon](t)\, \varepsilon_t(t) \quad \text{a.e. in } (0, T) \tag{1.21}$$

for all $(\sigma^0_r, \varepsilon) \in [-r, r] \times W^{1,1}(0, T)$, and the difference between the right and the left of (1.21) is the dissipated energy. Equation (1.21) can also be interpreted as a *chain rule inequality* for the energy operator $\mathcal{P}_r$ where the stop operator s_r plays the role of the "derivative" of $\mathcal{P}_r$ with respect to ε (only formally, since $\mathcal{P}_r$ is certainly not differentiable with respect to ε).

Chain rule inequalities of the form (1.21) have proven to be crucial for a successful study of differential equations with hysteresis (for this, see the cited literature). In the case of system (1.1)–(1.7), an appropriate form of such a condition is to postulate the existence of a further hysteresis operator $\mathcal{F}_2$ such that for any $(\varepsilon, w) \in W^{1,1}(0,T) \times W^{1,1}(0,T)$ it holds, for a.e. $t \in (0,T)$, that

$$\begin{aligned} &\frac{d}{dt}\,\mathcal{F}_1[\varepsilon,w](t) \le \mathcal{H}_1[\varepsilon,w](t)\,\varepsilon_t(t) + \mathcal{H}_3[\varepsilon,w](t)\,w_t(t), \\ &\frac{d}{dt}\,\mathcal{F}_2[\varepsilon,w](t) \le \mathcal{H}_2[\varepsilon,w](t)\,\varepsilon_t(t) + \mathcal{H}_4[\varepsilon,w](t)\,w_t(t). \end{aligned} \tag{1.22}$$

We then can associate with system (1.1)–(1.7) free energy, entropy, and internal energy *hysteresis operators* by putting (compare (1.8), (1.10))

$$\begin{aligned} &\mathcal{F}[\varepsilon,w,\theta] := \theta(1-\log(\theta)) + \mathcal{F}_1[\varepsilon,w] + \theta\,\mathcal{F}_2[\varepsilon,w], \\ &\mathcal{S}[\varepsilon,w,\theta] := \log(\theta) - \mathcal{F}_2[\varepsilon,w], \\ &\mathcal{U}[\varepsilon,w,\theta] := \theta + \mathcal{F}_1[\varepsilon,w], \end{aligned} \tag{1.23}$$

where $[\varepsilon, w, \theta] \in \mathrm{Map}[0,T] \times \mathrm{Map}[0,T] \times (0,+\infty)$. Indeed, if we associate σ and ψ as given by (1.2) and (1.5), respectively, with the "derivatives" of $\mathcal{F}$ with respect to ε and w (only formally, as they do not exist), respectively, then we arrive at system (1.1)–(1.5) as field equations. It will turn out later that the validity of (1.22) (rather, of a generalized version thereof; see below) will guarantee the thermodynamic consistency of the model, that is, the temperature stays positive during the evolution, and the *Clausius–Duhem inequality*, which in view of (1.12) can be written in the form

$$\theta\,\frac{d}{dt}\,\mathcal{S}[\varepsilon,w,\theta] - \frac{d}{dt}\,\mathcal{U}[\varepsilon,w,\theta] \ge -\tilde{\sigma}\varepsilon_t \quad \text{a.e. in } \Omega_T, \tag{1.24}$$

where $\tilde{\sigma} = \sigma + \varepsilon_t$ again denotes the total stress, will be satisfied.

The rest of the paper is organized as follows: In section 2, we give a detailed statement of the mathematical problem and of the main mathematical result. Section 3 brings the proof of local existence and global uniqueness, and in the concluding section 4 we prove global existence for system (1.1)–(1.7).

In what follows, the norms of the standard Lebesgue spaces $L^p(\Omega)$ for $1 \le p \le \infty$ will be denoted by $\|\cdot\|_p$. Finally, we shall use the usual denotations $W^{m,p}(\Omega)$ and $H^m(\Omega)$, $m \in \mathbb{N}$, $1 \le p \le \infty$, for the standard Sobolev spaces.

2. Statement of the problem. We make the following general assumptions on the data of the system.

(H1) $u_0 \in H^2(\Omega)$, $u_1 \in H^1(\Omega)$, $\theta_0 \in H^1(\Omega)$, $w_0 \in H^1(\Omega)$, it holds that $\theta_0(x) \ge \delta > 0$ for all $x \in \bar{\Omega}$, and the compatibility condition $u_0(0) = u_1(0) = 0$ is satisfied.

(H2) It holds that $f \in H^1(0,T;L^2(\Omega))$.

(H3) We assume that $g : \Omega \times (0,T) \times \mathbb{R} \to \mathbb{R}$ is a Carathéodory function such that

$$\exists\, g_0 \in L^\infty(\Omega_T) : \quad \theta \le 0 \;\Rightarrow\; g(x,t,\theta) = g_0(x,t), \tag{2.1}$$

$$\begin{aligned} &\exists\, K_1 > 0 : \quad |g(x,t,\theta_1) - g(x,t,\theta_2)| \le K_1|\theta_1 - \theta_2| \quad \text{for a.e.} \\ &(x,t) \in \Omega \times (0,T) \text{ and } \forall \theta_1, \theta_2 \in \mathbb{R}, \end{aligned} \tag{2.2}$$

$$g_0(x,t) \geq 0 \qquad \text{a.e. in } \Omega_T. \tag{2.3}$$

(H4) The operators $\mathcal{H}_j$, $1 \leq j \leq 4$, and $\mathcal{F}_1$ are causal and map $C[0,T] \times C[0,T]$ continuously into $C[0,T]$, and $W^{1,1}(0,T) \times W^{1,1}(0,T)$ into $W^{1,1}(0,T)$. In addition, the following conditions are satisfied:

(i) $\exists\, K_2 > 0 :\ \forall\, \varepsilon, w \in C[0,T]$ it holds that

$$\max_{j\in\{2,4\}} \|\mathcal{H}_j[\varepsilon,w]\|_\infty \leq K_2, \quad \mathcal{F}_1[\varepsilon,w](t) \geq -K_2 \quad \forall\, t \in [0,T]. \tag{2.4}$$

(ii) $\exists\, K_3 > 0 :\ \forall\, \varepsilon, w \in W^{1,1}(0,T)$ it holds, for a.e. $t \in (0,T)$, that

$$\max_{1\leq j\leq 4} |\mathcal{H}_j[\varepsilon,w]_t(t)| + |\mathcal{F}_1[\varepsilon,w]_t(t)| \leq K_3\Big(|\varepsilon_t(t)| + |w_t(t)|\Big). \tag{2.5}$$

(iii) $\exists\, K_4 > 0 :\ \forall\, \varepsilon_1, w_1, \varepsilon_2, w_2 \in C[0,T]$ it holds, for every $t \in [0,T]$, that

$$\begin{aligned} &\max_{1\leq j\leq 4} |\mathcal{H}_j[\varepsilon_1,w_1](t) - \mathcal{H}_j[\varepsilon_2,w_2](t)| \\ &\leq K_4 \max_{0\leq r\leq t}\Big(|\varepsilon_1(r) - \varepsilon_2(r)| + |w_1(r) - w_2(r)|\Big), \end{aligned} \tag{2.6}$$

and $\forall\, \varepsilon_1, w_1, \varepsilon_2, w_2 \in W^{1,1}(0,T)$ it holds, for every $t \in [0,T]$, that

$$\begin{aligned} |\mathcal{F}_1[\varepsilon_1,w_1](t) - \mathcal{F}_1[\varepsilon_2,w_2](t)| \leq K_4\Big[&|\varepsilon_1(0) - \varepsilon_2(0)| + |w_1(0) - w_2(0)| \\ &+ \int_0^t \Big(|\varepsilon_{1,t}(r) - \varepsilon_{2,t}(r)| + |w_{1,t}(r) - w_{2,t}(r)|\Big)\, dr\Big]. \end{aligned} \tag{2.7}$$

(H5) There exist causal operators $\mathcal{F}_2 : W^{1,1}(0,T) \times W^{1,1}(0,T) \to W^{1,1}(0,T)$, $\mathcal{G} : W^{1,1}(0,T) \to W^{1,1}(0,T)$, and a constant $K_5 > 0$ such that the following conditions are satisfied:

(i) For every $\varepsilon, w \in W^{1,1}(0,T)$ it holds that

$$\mathcal{F}_1[\varepsilon,w]_t \leq \varepsilon_t\, \mathcal{H}_1[\varepsilon,w] + \mathcal{G}[w]_t\, \mathcal{H}_3[\varepsilon,w] \quad \text{a.e. in } (0,T), \tag{2.8}$$

$$\mathcal{F}_2[\varepsilon,w]_t \leq \varepsilon_t\, \mathcal{H}_2[\varepsilon,w] + \mathcal{G}[w]_t\, \mathcal{H}_4[\varepsilon,w] \quad \text{a.e. in } (0,T). \tag{2.9}$$

(ii) For every $w \in W^{1,1}(0,T)$ it holds that

$$|\mathcal{G}[w]_t(t)|^2 \leq K_5 w_t(t)\, \mathcal{G}[w]_t(t) \quad \text{for a.e. } t \in (0,T). \tag{2.10}$$

Remark 1. Owing to (H4)(iii) we have, in particular, that for any $\varepsilon, w \in H^1(0,T)$ and $t \in [0,T]$ the following holds:

$$\begin{aligned} &|\mathcal{H}_1[\varepsilon,w](t)| + |\mathcal{H}_3[\varepsilon,w](t)| \\ &\leq |\mathcal{H}_1[\varepsilon,w](0)| + |\mathcal{H}_3[\varepsilon,w](0)| + 2K_4 \max_{0\leq r\leq t} (|\varepsilon(r) - \varepsilon(0)| + |w(r) - w(0)|) \\ &\leq |\mathcal{H}_1[\varepsilon,w](0)| + |\mathcal{H}_3[\varepsilon,w](0)| + 2K_4 \int_0^t \Big(|\varepsilon_t(r)| + |w_t(r)|\Big)\, dr \\ &\leq |\mathcal{H}_1[\varepsilon,w](0)| + |\mathcal{H}_3[\varepsilon,w](0)| + 2K_4\sqrt{t} \left(\int_0^t \Big(|\varepsilon_t(r)|^2 + |w_t(r)|^2\Big)\, dr\right)^{1/2}. \end{aligned} \tag{2.11}$$

In addition, a linear growth of $\mathcal{H}_1$ and $\mathcal{H}_3$ with respect to both ε and w is admitted, which, in particular, includes the case of simple linear elasticity. It also follows that for any $\varepsilon, w \in H^1(\Omega; C[0,T])$ it holds, for a.e. $(x,t) \in \Omega_T$, that

$$\max_{1\le j\le 4} \Big| \Big(\mathcal{H}_j[\varepsilon, w]\Big)_x (x,t)\Big| \le K_4 \max_{0\le r\le t} \Big(|\varepsilon_x(x,r)| + |w_x(x,r)|\Big). \tag{2.12}$$

Indeed, we only have to apply (2.6) with $\varepsilon_1(x,t) := \varepsilon(x+h,t)$, $\varepsilon_2(x,t) := \varepsilon(x,t)$, $w_1(x,t) := w(x+h,t)$, $w_2(x,t) := w(x,t)$, with some $h > 0$, and then let $h \searrow 0$. Consequently, we may consider first order spatial derivatives of $\mathcal{H}_j[\varepsilon, w]$, and we have $(\mathcal{H}_j[\varepsilon, w])_x \in L^2(\Omega_T)$, $1 \le j \le 4$.

Remark 2. A typical example where (H4), (H5) are fulfilled is given by *Prandtl–Ishlinskii operators* of the form

$$\begin{aligned} \mathcal{H}_j[\varepsilon, w] &:= \int_0^\infty \varphi_j(r)\, s_r\left[\sigma_r^{0,j}, \varepsilon\right]\, dr, \quad j = 1,2, \\ \mathcal{H}_j[\varepsilon, w] &:= \int_0^\infty \varphi_j(r)\, s_r\left[\sigma_r^{0,j}, w\right]\, dr, \quad j = 3,4, \end{aligned} \tag{2.13}$$

where $\sigma_r^{0,j} \in [-r, +r]$, $1 \le j \le 4$, are given initial values for the operators s_r defined in (1.17), and the weight functions φ_j are nonnegative on $[0, +\infty)$ and satisfy

$$\max_{1\le j\le 4} \int_0^\infty (1+r^2)\, \varphi_j(r)\, dr < +\infty. \tag{2.14}$$

Indeed, defining the (energy) operators

$$\begin{aligned} \mathcal{F}_1[\varepsilon, w] &:= \frac{1}{2}\int_0^\infty \Big(\varphi_1(r)\, s_r^2\left[\sigma_r^{0,1}, \varepsilon\right] + \varphi_3(r)\, s_r^2\left[\sigma_r^{0,3}, w\right]\Big)\, dr, \\ \mathcal{F}_2[\varepsilon, w] &:= \frac{1}{2}\int_0^\infty \Big(\varphi_2(r)\, s_r^2\left[\sigma_r^{0,2}, \varepsilon\right] + \varphi_4(r)\, s_r^2\left[\sigma_r^{0,4}, w\right]\Big)\, dr, \end{aligned} \tag{2.15}$$

choosing $\mathcal{G}[w] = w$, and invoking the properties (1.21)–(1.24) of the stop operators s_r, we easily verify the validity of (H4), (H5). Other examples, where the dependence on ε, w is no longer decoupled as in (2.13), can be constructed using multidimensional stop operators as basic elements (cf. [15], [16]). For examples where the $\mathcal{H}_j$ are not Prandtl–Ishlinskii operators and $\mathcal{G}$ differs from the identity operator, we refer to [14], [15].

We can now formulate the main result of this paper.

THEOREM 2.1. *Suppose that the hypotheses* (H1) *to* (H5) *are satisfied. Then the system* (1.1)–(1.7) *admits a unique strong solution* (u, θ, w) *such that* (1.1)–(1.5) *hold a.e. in* Ω_T*, and such that*

$$\begin{aligned} &u \in H^2(0,T; L^2(\Omega)) \cap H^1(0,T; H^2(\Omega)),\ w \in H^2(0,T; L^2(\Omega)) \cap H^1(0,T; H^1(\Omega)), \\ &\theta \in H^1(0,T; L^2(\Omega)) \cap L^2(0,T; H^2(\Omega)). \end{aligned} \tag{2.16}$$

In addition, with the finite norms $\beta_1 := \|u_{xt}\|_{L^1(0,T;L^\infty(\Omega))}$ *and* $\beta_2 := \|w_t\|_{C(\overline{\Omega_T})}$ *it holds that*

$$\theta(x,t) \ge \delta e^{-((K_1 + K_2K_5\beta_2)t + K_2\beta_1)} \quad \forall (x,t) \in \overline{\Omega_T}. \tag{2.17}$$

Remark 3. We note at this point that Theorem 2.1 also implies that the second principle of thermodynamics is satisfied for the system (1.1)–(1.7). Indeed, we have

$\theta > 0$ in $\overline{\Omega_T}$, and the validity of the Clausius–Duhem inequality (1.24) follows from the simple calculation

$$\begin{aligned} &\theta\,\mathcal{S}[\varepsilon,w,\theta]_t - \mathcal{U}[\varepsilon,w,\theta]_t + \tilde{\sigma}\varepsilon_t = -\theta\,\mathcal{F}_2[\varepsilon,w]_t - \mathcal{F}_1[\varepsilon,w]_t + \sigma\varepsilon_t + \varepsilon_t^2 \\ &\geq -\Big(\mathcal{H}_1[\varepsilon,w] + \theta\,\mathcal{H}_2[\varepsilon,w]\Big)\varepsilon_t - \Big(\mathcal{H}_3[\varepsilon,w] + \theta\,\mathcal{H}_4[\varepsilon,w]\Big)\mathcal{G}[w]_t + \sigma\varepsilon_t + \varepsilon_t^2 \\ &\geq \varepsilon_t^2 + w_t\,\mathcal{G}[w]_t \geq 0 \quad \text{a.e. in } \Omega_T, \end{aligned} \tag{2.18}$$

where $\mathcal{F}, \mathcal{S}, \mathcal{U}$ are given by (1.23). We may therefore claim that our system is thermodynamically consistent.

The proof of Theorem 2.1 will be given in the following sections. During its course, we will make repeated use of *Young's inequality*,

$$ab \leq \frac{\delta}{2}a^2 + \frac{1}{2\delta}b^2 \quad \forall\, a,b \in \mathbb{R},\ \delta > 0, \tag{2.19}$$

of the elementary inequality,

$$|z(t)|^2 \leq 2|z(0)|^2 + 2t\int_0^t z_t^2(r)\,dr \quad \forall\, t \in (0,T) \quad \forall\, z \in H^1(0,T), \tag{2.20}$$

and of the one-dimensional *Gagliardo–Nirenberg inequality*,

$$\|w\|_p \leq K_0\Big(\|w\|_q^{1-\omega}\|w_x\|_r^{\omega} + \|w\|_q\Big) \quad \forall\, w \in W^{1,r}(\Omega) \cap L^q(\Omega), \tag{2.21}$$

where $K_0 > 0$ is a constant depending only on p, q, r, and where

$$1 \leq r \leq +\infty, \quad 1 \leq q \leq p \leq +\infty, \quad \omega\left(\frac{1}{q} - \frac{1}{r} + 1\right) = \frac{1}{q} - \frac{1}{p}. \tag{2.22}$$

3. Local existence. We rewrite the system (1.1)–(1.7), using the transformation due to Andrews [1],

$$u_x = p + q, \quad \text{where } p(x,t) := \int_1^x u_t(\xi,t)\,d\xi. \tag{3.1}$$

We easily find that (1.1)–(1.6) is equivalent to the system

$$p_t - p_{xx} = \sigma + \int_1^x f(\xi,t)\,d\xi, \tag{3.2}$$

$$p(1,t) = p_x(0,t) = 0, \quad p(x,0) = \int_1^x u_1(\xi)\,d\xi, \tag{3.3}$$

$$\sigma = \mathcal{H}_1[p+q,w] + \theta\,\mathcal{H}_2[p+q,w], \tag{3.4}$$

$$q_t = -\sigma - \int_1^x f(\xi,t)\,d\xi, \tag{3.5}$$

$$q(x,0) = u_0'(x) - \int_1^x u_1(\xi)\,d\xi, \tag{3.6}$$

$$\Big(\theta + \mathcal{F}_1[p+q,w]\Big)_t - \theta_{xx} = p_{xx}^2 + \sigma p_{xx} + g(x,t,\theta), \tag{3.7}$$

$$w_t = -\psi, \tag{3.8}$$

$$\psi = \mathcal{H}_3[p+q,w] + \theta\,\mathcal{H}_4[p+q,w], \tag{3.9}$$

$$\theta(x,0) = \theta_0(x), \quad w(x,0) = w_0(x), \quad \theta_x(0,t) = \theta_x(1,t) = 0. \tag{3.10}$$

Let $V_0 := \{z \in H^1(\Omega); z(1) = 0\}$, and let V_0^* denote its dual space. We are going to show the following result.

THEOREM 3.1. *Suppose that the hypotheses* (H1) *to* (H4) *are fulfilled. Then there is some* $\hat\tau > 0$ *such that the initial-boundary value problem* (3.2)–(3.10) *admits a unique solution quadruple* (p,q,θ,w) *on* $\bar\Omega \times [0,\hat\tau]$ *satisfying*

$$p \in H^2(0,\hat\tau;V_0^*) \cap H^1(0,\hat\tau;H^1(\Omega)) \cap L^2(0,\hat\tau;H^3(\Omega)), \tag{3.11}$$

$$q, w \in H^2(0,\hat\tau;L^2(\Omega)) \cap H^1(0,\hat\tau;H^1(\Omega)) \cap C^1([0,\hat\tau];C(\bar\Omega)), \tag{3.12}$$

$$\theta \in H^1(0,\hat\tau;L^2(\Omega)) \cap L^2(0,\hat\tau;H^2(\Omega)) \cap C(\overline{\Omega_{\hat\tau}}), \tag{3.13}$$

$$\theta(x,t) \ge \frac{\delta}{2} > 0 \quad \textit{for every } (x,t) \in \bar\Omega \times [0,\hat\tau]. \tag{3.14}$$

Proof of Theorem 3.1. We divide the proof of Theorem 3.1 into several steps, each formulated as a separate lemma. The existence part of the proof is based on the following special case of the *Schauder–Tikhonov fixed point principle* (cf., for instance, Theorem 3.6.1 in [6]).

LEMMA 3.2. *Let the operator* $\mathcal{T}$ *map the nonempty, closed, convex, and weakly compact subset* $\mathcal{M}$ *of the separable Hilbert space* X *into itself, and suppose that* $\mathcal{T}$ *is weakly sequentially continuous on* $\mathcal{M}$*; that is, it holds that* $\mathcal{T}(v_n) \to \mathcal{T}(v)$ *weakly in* X *whenever* $v_n \to v$ *weakly in* X *for some sequence* $\{v_n\} \subset \mathcal{M}$*. Then* $\mathcal{T}$ *has a fixed point in* $\mathcal{M}$*.*

We aim to apply Lemma 3.2 to the following setting. Consider for $\tau \in (0,T]$ the separable Hilbert spaces

$$\begin{aligned} P_\tau &:= H^2(0,\tau;V_0^*) \cap H^1(0,\tau;H^1(\Omega)) \cap L^2(0,\tau;H^3(\Omega)),\\ Q_\tau &:= H^2(0,\tau;L^2(\Omega)) \cap H^1(0,\tau;H^1(\Omega)),\\ Z_\tau &:= H^1(0,\tau;L^2(\Omega)) \cap L^2(0,\tau;H^2(\Omega)),\\ W_\tau &:= H^2(0,\tau;L^2(\Omega)) \cap H^1(0,\tau;H^1(\Omega)),\\ X_\tau &:= P_\tau \times Q_\tau \times Z_\tau \times W_\tau \end{aligned} \tag{3.15}$$

and introduce the sets

$$\mathcal{M}_\tau := \Big\{ (p,q,\theta,w) \in X_\tau;\ (3.3),\ (3.6),\ (3.10) \text{ hold},\ p_t + q_t = p_{xx} \quad \text{a.e. in } \Omega_\tau,$$

$$\int_0^\tau \int_\Omega \Big(\theta_t^2 + \theta_{xx}^2\Big)\,dx\,dt + \max_{0\le t\le\tau} \int_\Omega |\theta_x(x,t)|^2\,dx \le M_1, \tag{3.16}$$

$$\max_{(x,t)\in\overline{\Omega_\tau}} |\theta(x,t)| \le M_2, \tag{3.17}$$

$$\max_{0\le t\le\tau} \int_\Omega \Big(|q_t(x,t)|^2 + |w_t(x,t)|^2\Big)\,dx \le M_3, \tag{3.18}$$

$$\int_0^\tau \int_\Omega \Big(q_{xt}^2 + w_{xt}^2 \Big)\, dx\, dt \le M_4, \tag{3.19}$$

$$\int_0^\tau \int_\Omega \Big(p_t^2 + p_{xx}^2 \Big)\, dt + \max_{0\le t\le\tau} \int_\Omega |p_x(x,t)|^2\, dx \le M_5, \tag{3.20}$$

$$\max_{0\le t\le\tau} \int_\Omega \Big(|q_x(x,t)|^2 + |w_x(x,t)|^2 \Big)\, dx \le M_6, \tag{3.21}$$

$$\|p_t\|^2_{H^1(0,\tau;V_0^*)} + \int_0^\tau \int_\Omega p_{xt}^2\, dx\, dt + \max_{0\le t\le\tau} \int_\Omega |p_t(x,t)|^2\, dx \le M_7, \tag{3.22}$$

$$\max_{0\le t\le\tau} \int_\Omega |p_{xx}(x,t)|^2\, dx \le M_8, \tag{3.23}$$

$$\int_0^\tau \int_\Omega p_{xxx}^2\, dx\, dt \le M_9, \tag{3.24}$$

$$\int_0^\tau \int_\Omega (q_{tt}^2 + w_{tt}^2)\, dx\, dt \le M_{10}, \tag{3.25}$$

$$\min_{(x,t)\in\overline{\Omega_\tau}} \theta(x,t) \ge \frac{\delta}{2} > 0 \Big\}, \tag{3.26}$$

where the positive constants M_i, $i = 1, \dots, 10$, will have to be specified later. Obviously, $\mathcal{M}_\tau$ is a nonempty, closed, convex, and bounded (hence weakly compact) subset of the separable Hilbert space X_τ.

Next, we introduce the operator $\mathcal{T}$ on $\mathcal{M}_\tau$ by $\mathcal{T}(\bar p, \bar q, \bar\theta, \bar w) := (p, q, \theta, w)$, where for $(\bar p, \bar q, \bar\theta, \bar w) \in \mathcal{M}_\tau$ the quadruple (p, q, θ, w) is the unique solution to the linear initial-boundary value problem

$$p_t - p_{xx} = \bar\sigma + \int_1^x f(\xi, t)\, d\xi, \tag{3.27}$$

$$p(1,t) = p_x(0,t) = 0, \quad p(x,0) = \int_1^x u_1(\xi)\, d\xi, \tag{3.28}$$

$$\bar\sigma = \mathcal{H}_1[\bar p + \bar q, \bar w] + \bar\theta\, \mathcal{H}_2[\bar p + \bar q, \bar w], \tag{3.29}$$

$$q_t = -\bar\sigma - \int_1^x f(\xi, t)\, d\xi, \tag{3.30}$$

$$q(x,0) = u_0'(x) - \int_1^x u_1(\xi)\, d\xi, \tag{3.31}$$

$$\theta_t - \theta_{xx} = -\mathcal{F}_1[\bar p + \bar q, \bar w]_t + \bar p_{xx}^2 + \bar\sigma \bar p_{xx} + g(x, t, \bar\theta), \tag{3.32}$$

$$w_t = -\bar\psi, \tag{3.33}$$

$$\bar\psi = \mathcal{H}_3[\bar p + \bar q, \bar w] + \bar\theta\, \mathcal{H}_4[\bar p + \bar q, \bar w], \tag{3.34}$$

$$\theta(x,0) = \theta_0(x), \quad w(x,0) = w_0(x), \quad \theta_x(0,t) = \theta_x(1,t) = 0. \tag{3.35}$$

We have the following result.

LEMMA 3.3. *There exist* $\hat\tau \in (0,T]$ *and positive constants* M_i, $i = 1,\dots,10$, *such that* $\mathcal{T}(\mathcal{M}_\tau) \subset \mathcal{M}_\tau$ *for any* $\tau \in (0,\hat\tau]$.

Proof. Let $\tau \in (0,T]$ be given. Without loss of generality, we may assume that $\tau \le 1$. We have $p_t + q_t = p_{xx}$ a.e. in Ω_τ, and we infer from the general hypotheses that $p_t, p_{xx}, q_t, w_t, \theta_t, \theta_{xx} \in L^2(\Omega_\tau)$. Therefore, $\theta \in Z_\tau$. Also, since $(\bar p, \bar q, \bar\theta, \bar w) \in \mathcal{M}_\tau$, it follows from Remark 1 that the right-hand sides of both (3.30) and (3.33) belong to $H^1(\Omega_\tau)$ so that $q \in Q_\tau$ and $w \in W_\tau$.

Next, we consider the parabolic initial-boundary value problem

$$z_t - z_{xx} = v := \bar\sigma_t + \int_1^x f_t(\xi,t)\,d\xi, \tag{3.36}$$

$$z_x(0,t) = z(1,t) = 0, \quad z(x,0) = u_1'(x) + \bar\sigma(0) + \int_1^x f(\xi,0)\,d\xi. \tag{3.37}$$

Since $z(\cdot,0) \in L^2(\Omega)$, and since the right-hand side v of (3.36) belongs to $L^2(\Omega_\tau)$, it follows from general linear parabolic theory (cf. Lions and Magenes [19]) that (3.36)–(3.37) admits a unique weak solution $z \in L^2(0,\tau;H^1(\Omega)) \cap H^1(0,\tau;V_0^*) \cap C([0,\tau];L^2(\Omega))$, and there is some constant $\hat C > 0$, not depending on $\tau \in (0,T]$, such that

$$\begin{aligned} &\|z\|^2_{H^1(0,\tau;V_0^*)} + \int_0^\tau\int_\Omega z_x^2\,dx\,dt + \max_{0\le t\le\tau}\int_\Omega |z(x,t)|^2\,dx \\ &\le \hat C\left(\int_\Omega |z(x,0)|^2\,dx + \int_0^\tau\int_\Omega v^2\,dx\,dt\right). \end{aligned} \tag{3.38}$$

Invoking the compatibility condition $u_1(0) = 0$, we easily verify that

$$p(x,t) = \int_1^x u_1(\xi)\,d\xi + \int_0^t z(x,r)\,dr, \tag{3.39}$$

so that $p \in H^2(0,\tau;V_0^*) \cap H^1(0,\tau;H^1(\Omega)) \cap C^1([0,\tau];L^2(\Omega))$, and (3.38) holds for $z = p_t$. Hence, using (3.27), we can conclude that $p_{xxx} \in L^2(\Omega_\tau)$, and also $p_{xx} \in C([0,\tau];L^2(\Omega))$. In conclusion, $p \in P_\tau$, and we have shown that $\mathcal{T}(\mathcal{M}_\tau) \subset X_\tau$.

Now let $(p,q,\theta,w) = \mathcal{T}(\bar p,\bar q,\bar\theta,\bar w)$ for some $(\bar p,\bar q,\bar\theta,\bar w) \in \mathcal{M}_\tau$, where the constants $M_1,\dots,M_{10}$ and $\tau \in (0,T]$ are assumed to be fixed. We are going to derive a number of estimates for (p,q,θ,w) in terms of $M_1,\dots,M_{10}$ and of the data of the system. In what follows, we denote by C_i, $i \in \mathbb{N} \cup \{0\}$, positive constants which may depend on the given data u_0, u_1, θ_0, w_0, f, g_0, and on the constants K_i, $0 \le i \le 4$, but neither on τ nor on $M_1,\dots,M_{10}$.

At first, we conclude from (2.11), (2.4), and (2.5) that

$$\begin{aligned} &\int_\Omega \Big(|\bar\sigma|^2 + |\bar\psi|^2\Big)(x,t)\,dx \\ &\le 2\int_\Omega \Big(\mathcal{H}_1^2[\bar p + \bar q, \bar w] + \mathcal{H}_3^2[\bar p + \bar q, \bar w] + \bar\theta^2(\mathcal{H}_2^2[\bar p + \bar q, \bar w] + \mathcal{H}_4^2[\bar p + \bar q, \bar w])\Big)(x,t)\,dx \\ &\le 4K_2^2M_2^2 + C_0\left[1 + t\int_0^t\int_\Omega \Big(\bar p_t^2 + \bar q_t^2 + \bar w_t^2\Big)\,dx\,dr\right] \\ &\le 4K_2^2M_2^2 + C_0\Big(1 + t(M_3 + M_7)\Big) \quad \forall t \in [0,\tau], \end{aligned} \tag{3.40}$$

$$|\bar\sigma_t| + |\bar{\psi}_t| \le 2K_3(1+M_2)(|\bar p_t| + |\bar q_t| + |\bar w_t|) + 2K_2|\bar\theta_t| \quad \text{a.e. in } \Omega_\tau. \tag{3.41}$$

Also, it follows from (2.12) that for $1 \le i \le 4$ and a.e. $(x,t) \in \Omega_\tau$ we have

$$\Big|\Big(\mathcal{H}_i[\bar p + \bar q, \bar w]\Big)_x(x,t)\Big| \le K_4 \max_{0\le r\le t}\Big(|\bar p_x(x,r)| + |\bar q_x(x,r)| + |\bar w_x(x,r)|\Big). \tag{3.42}$$

In addition, owing to (2.5),

$$|\mathcal{F}_1[\bar p + \bar q, \bar w]_t| \le K_3(|\bar p_t| + |\bar q_t| + |\bar w_t|) \quad \text{a.e. in } \Omega_\tau, \tag{3.43}$$

as well as, by virtue of (H3),

$$|g(x,t,\bar\theta(x,t))| \le g_0(x,t) + K_1 M_2. \tag{3.44}$$

Now multiply (3.32) first by θ_t and then by $-\theta_{xx}$; add the resulting equations and integrate over $\Omega\times[0,t]$ for any $t\in(0,\tau]$. Using Young's inequality and invoking (3.40), (3.43), and (3.44), we find that

$$\begin{aligned} &\int_0^t\int_\Omega \Big(\theta_t^2 + \theta_{xx}^2\Big)\,dx\,dr + \int_\Omega |\theta_x(x,t)|^2\,dx \\ &\le C_1\left[1 + tM_2^2 + \int_0^t\int_\Omega\Big(\bar p_t^2 + \bar q_t^2 + \bar w_t^2\Big)\,dx\,dr \right. \\ &\qquad \left. + \int_0^t\int_\Omega \bar p_{xx}^4\,dx\,dr + \int_0^t\int_\Omega \bar\sigma^2\,\bar p_{xx}^2\,dx\,dr\right]. \end{aligned} \tag{3.45}$$

Invoking the Gagliardo–Nirenberg inequality (2.21) for $p=+\infty$, $q=r=2$, $\omega=1/2$, we infer that

$$\begin{aligned} \int_0^t \|\bar p_{xx}(\cdot,r)\|_\infty^2\,dr &\le 2K_0^2\left(\int_0^t\int_\Omega \bar p_{xx}^2\,dx\,dr \right. \\ &\qquad \left. + \max_{0\le r\le t}\|\bar p_{xx}(\cdot,r)\|_2 \int_0^t \|\bar p_{xxx}(\cdot,r)\|_2\,dr\right) \\ &\le 2K_0^2\Big(tM_8 + \sqrt{M_8}\sqrt{t}\sqrt{M_9}\Big) \le 2K_0^2\sqrt{t}\Big(M_8 + \sqrt{M_8M_9}\Big). \end{aligned} \tag{3.46}$$

It follows that

$$\begin{aligned} \int_0^t\int_\Omega \bar p_{xx}^4\,dx\,dr &\le \max_{0\le r\le t}\|\bar p_{xx}(\cdot,r)\|_2^2 \int_0^t \|\bar p_{xx}(\cdot,r)\|_\infty^2\,dr \\ &\le 2K_0^2\sqrt{t}\Big(M_8^2 + M_8^{3/2}\sqrt{M_9}\Big). \end{aligned} \tag{3.47}$$

In addition, by (3.40),

$$\begin{aligned} &\int_0^t\int_\Omega \bar\sigma^2\bar p_{xx}^2\,dx\,dr \le \max_{0\le r\le t}\|\bar\sigma(\cdot,r)\|_2^2 \int_0^t \|\bar p_{xx}(\cdot,r)\|_\infty^2\,dr \\ &\le 2K_0^2\sqrt{t}\Big(M_8 + \sqrt{M_8M_9}\Big)\Big(4K_2^2M_2^2 + C_0(1+t(M_3+M_7))\Big). \end{aligned} \tag{3.48}$$

In conclusion, we have shown the estimate

$$\begin{aligned} &\int_0^\tau\int_\Omega \Big(\theta_t^2 + \theta_{xx}^2\Big)\,dx\,dr + \max_{0\le t\le\tau}\int_\Omega |\theta_x(x,t)|^2\,dx \\ &\le C_2\Big[1 + \sqrt{\tau}\big(M_2^2 + M_3 + M_7 + M_8^2 + M_8^{3/2}M_9^{1/2} \\ &\qquad + \big(M_8 + M_8^{1/2}M_9^{1/2}\big)\big(1 + M_2^2 + M_3 + M_7\big)\big)\Big]. \end{aligned} \tag{3.49}$$

Next, we consider (3.30) and (3.33). By the general hypotheses and (3.40), we have

$$\max_{0\le t\le\tau}\int_\Omega \Big(|q_t(x,t)|^2+|w_t(x,t)|^2\Big)\,dx \le C_3\Big(1+M_2^2+t(M_3+M_7)\Big). \tag{3.50}$$

Now differentiate (3.30) and (3.33) with respect to x. Then, by (3.42),

$$|\bar\sigma_x(x,t)|+|\bar\psi_x(x,t)| \le 2K_2|\bar\theta_x(x,t)|+2K_4(1+M_2)\max_{0\le r\le t}\Big((|\bar p_x|+|\bar q_x|+|\bar w_x|)(x,r)\Big) \tag{3.51}$$

for a.e. $(x,t)\in\Omega_\tau$. Therefore, using (2.20), we can conclude that

$$\begin{aligned}&\int_0^\tau\int_\Omega\Big(q_{xt}^2+w_{xt}^2\Big)\,dx\,dt \le C_4\Big(1+\int_0^\tau\int_\Omega\Big(|\bar\sigma_x|^2+|\bar\psi_x|^2\Big)\,dx\,dt\Big)\\ &\le C_5\left[1+M_2^2+\int_0^\tau\int_\Omega|\bar\theta_x|^2\,dx\,dt+(1+M_2^2)\tau\int_0^\tau\int_\Omega\Big(\bar p_{xt}^2+\bar q_{xt}^2+\bar w_{xt}^2\Big)\,dx\,dt\right]\\ &\le C_5\Big(1+M_2^2+\tau\big(M_1+(1+M_2^2)(M_4+M_7)\big)\Big).\end{aligned} \tag{3.52}$$

But then also

$$\begin{aligned}&\max_{0\le t\le\tau}\int_\Omega\Big(|q_x(x,t)|^2+|w_x(x,t)|^2\Big)\,dx\\ &\le C_6\Big(1+M_2^2+\tau\left(M_1+(1+M_2^2)(M_4+M_7)\right)\Big).\end{aligned} \tag{3.53}$$

Next, we consider the linear parabolic system (3.27)–(3.28). Standard parabolic estimates, using the general hypotheses and (3.40), yield that

$$\begin{aligned}&\int_0^\tau\int_\Omega\Big(p_t^2+p_{xx}^2\Big)\,dx\,dt \le C_7\Big(1+\int_0^\tau\int_\Omega|\bar\sigma|^2\,dx\,dt\Big)\\ &\le C_8\Big(1+\tau(M_2^2+M_3+M_7)\Big).\end{aligned} \tag{3.54}$$

Moreover, since (3.38) is valid for $z=p_t$, we can infer from (H2) and from (3.41) that

$$\begin{aligned}&\|p\|^2_{H^2(0,\tau;V_0^*)}+\int_0^\tau\int_\Omega p_{xt}^2\,dx\,dt+\max_{0\le t\le\tau}\int_\Omega|p_t(x,t)|^2\,dx\\ &\le C_9\left(1+\int_0^\tau\int_\Omega\Big(\bar\theta_t^2+(1+M_2^2)\Big(\bar p_t^2+\bar q_t^2+\bar w_t^2\Big)\Big)\,dx\,dt\right)\\ &\le C_9\Big(1+M_1+\tau(1+M_2^2)(M_3+M_7)\Big).\end{aligned} \tag{3.55}$$

But then we obtain from (3.27), also using (3.29) and (H2), that

$$\begin{aligned}\max_{0\le t\le\tau}\int_\Omega|p_{xx}(x,t)|^2\,dx &\le 2\max_{0\le t\le\tau}\int_\Omega|p_t(x,t)|^2\,dx+2C_{10}\left(1+\max_{0\le t\le\tau}\int_\Omega|\bar\sigma(x,t)|^2\,dx\right)\\ &\le C_{11}\Big(1+M_1+M_2^2+\tau(1+M_2^2)(M_3+M_7)\Big).\end{aligned} \tag{3.56}$$

In addition, employing (3.51) and (3.55), and arguing as in the derivation of (3.52), we can deduce the estimate

$$\begin{aligned}\int_0^\tau\int_\Omega p_{xxx}^2\,dx\,dt &\le 2\int_0^\tau\int_\Omega p_{xt}^2\,dx\,dt+2\int_0^\tau\int_\Omega|\bar\sigma_x+f|^2\,dx\,dt\\ &\le C_{12}\Big(1+M_1+M_2^2+\tau(M_1+(1+M_2^2)(M_3+M_4+M_7))\Big).\end{aligned} \tag{3.57}$$

Next, we differentiate (3.30) and (3.33), respectively, with respect to t, and invoke (H2) and (3.41) to obtain the bound

$$\int_0^\tau \int_\Omega (q_{tt}^2 + w_{tt}^2)\,dx\,dt \le C_{13}(1 + (1 + M_2^2)(M_3 + M_5) + M_1). \tag{3.58}$$

Finally, since Ω is one-dimensional, $H^1(0,\tau;L^2(\Omega)) \cap L^2(0,\tau;H^2(\Omega))$ is continuously imbedded in $C(\overline{\Omega_\tau})$. Hence, there is some $C_{14} > 0$ such that

$$\max_{(x,t)\in\overline{\Omega_\tau}} |\theta(x,t)| \le C_{14}\left(1 + \sqrt{\hat{M}_1}\right), \tag{3.59}$$

where $\hat{M}_1$ is equal to the expression on the right-hand side of (3.49).

Now, we can define the constants $M_1 \dots, M_{10}$. We make the choices

$$\begin{aligned}
&M_1 := 2C_2, \quad M_2 := C_{14}\left(1 + \sqrt{\hat{M}_1}\right), \quad M_3 := 2C_3 + C_3 M_2^2,\\
&M_4 := 2C_5 + C_5 M_2^2, \quad M_5 := 2C_8, \quad M_6 := 2C_6 + C_6 M_2^2,\\
&M_7 := 2C_9 + C_9 M_1, \quad M_8 := 2C_{11} + C_{11}(M_1 + M_2^2),\\
&M_9 := 2C_{12} + C_{12}(M_1 + M_2^2),\ M_{10} := C_{13}(1 + (1 + M_2^2)(M_3 + M_5) + M_1).
\end{aligned} \tag{3.60}$$

It then follows from the estimates (3.49)–(3.50), (3.52)–(3.57), and (3.59) that there exists some $\tau_0 \in (0,T]$ such that the inequalities (3.16)–(3.27) are fulfilled for any $\tau \in (0,\tau_0]$. In particular, (3.16) implies that the assumptions of Lemma 3.2.2 in [2] are satisfied. Therefore, we have $\theta \in C^{\frac{1}{2},\frac{1}{6}}(\overline{\Omega_{\tau_0}})$, and there is some constant $C_{15} > 0$, depending only on M_1 and τ_0, such that for every $(x,t),(y,s) \in \overline{\Omega_{\tau_0}}$ it holds that

$$|\theta(x,t) - \theta(y,s)| \le C_{15}\left(|t-s|^{1/6} + |x-y|^{1/2}\right). \tag{3.61}$$

Consequently, for sufficiently small $\hat{\tau} \in (0,\tau_0]$,

$$\theta(x,t) \ge \theta_0(x) - |\theta(x,t) - \theta_0(x)| \ge \delta - C_{15} t^{1/6} \ge \frac{\delta}{2} \tag{3.62}$$

for all $(x,t) \in \Omega \times [0,\hat{\tau}]$. With this, the proof of the lemma is complete. ∎

LEMMA 3.4. *The operator $\mathcal{T}$ is weakly sequentially continuous in $\mathcal{M}_{\hat{\tau}}$.*

Proof. Suppose a sequence $\{(\bar{p}_n, \bar{q}_n, \bar{\theta}_n, \bar{w}_n)\} \subset \mathcal{M}_{\hat{\tau}}$ is given such that

$$(\bar{p}_n, \bar{q}_n, \bar{\theta}_n, \bar{w}_n) \to (\bar{p}, \bar{q}, \bar{\theta}, \bar{w}) \quad \text{weakly in } X_{\hat{\tau}} \quad \text{as } n \to \infty. \tag{3.63}$$

Since $\mathcal{M}_{\hat{\tau}}$ is weakly closed, it holds that $(\bar{p}, \bar{q}, \bar{\theta}, \bar{w}) \in \mathcal{M}_{\hat{\tau}}$. Now, let

$$(p_n, q_n, \theta_n, w_n) := \mathcal{T}(\bar{p}_n, \bar{q}_n, \bar{\theta}_n, \bar{w}_n), \quad n \in \mathbb{N}, \quad (p,q,\theta,w) := \mathcal{T}(\bar{p}, \bar{q}, \bar{\theta}, \bar{w}). \tag{3.64}$$

We have to show that

$$(p_n, q_n, \theta_n, w_n) \to (p, q, \theta, w) \quad \text{weakly in } X_{\hat{\tau}} \quad \text{as } n \to \infty. \tag{3.65}$$

Clearly, as $(p_n, q_n, \theta_n, w_n) \in \mathcal{M}_{\hat{\tau}}$ for all $n \in \mathbb{N}$, we have, on a subsequence which is again indexed by n,

$$(p_n, q_n, \theta_n, w_n) \to (\hat{p}, \hat{q}, \hat{\theta}, \hat{w}) \quad \text{weakly in } X_{\hat{\tau}} \quad \text{as } n \to \infty \tag{3.66}$$

for some $(\hat{p},\hat{q},\hat{\theta},\hat{w}) \in \mathcal{M}_{\hat{\tau}}$. It remains to show that $(\hat{p},\hat{q},\hat{\theta},\hat{w}) = (p,q,\theta,w)$. The uniqueness of the limit point then entails that (3.66), and thus (3.65) holds for the entire sequence. To this end, note that we have the convergences

$$\begin{aligned}&\bar{\theta}_{n,t} \to \bar{\theta}_t,\ \ \bar{\theta}_{n,xx} \to \bar{\theta}_{xx},\ \ \bar{p}_{n,t} \to \bar{p}_t,\ \ \bar{p}_{n,xx} \to \bar{p}_{xx},\ \ \bar{p}_{n,xxx} \to \bar{p}_{xxx},\ \ \bar{p}_{n,xt} \to \bar{p}_{xt},\\ &\bar{w}_{n,t} \to \bar{w}_t,\ \ \bar{w}_{n,xt} \to \bar{w}_{xt},\ \ \bar{q}_{n,t} \to \bar{q}_t,\ \ \bar{q}_{n,xt} \to \bar{q}_{xt},\ \text{ all weakly in } L^2(\Omega_{\hat{\tau}}).\end{aligned} \tag{3.67}$$

By compact imbedding, we may also assume that

$$\bar{\theta}_n \to \bar{\theta},\quad \bar{p}_{n,x} \to \bar{p}_x,\quad \text{both uniformly in } \overline{\Omega_{\hat{\tau}}}, \tag{3.68}$$

$$\bar{p}_n \to \bar{p},\quad \bar{q}_n \to \bar{q},\quad \bar{w}_n \to \bar{w},\quad \text{all strongly in } L^2(\Omega; C[0,\hat{\tau}]). \tag{3.69}$$

But then, owing to (H4), it follows that

$$\begin{aligned}&\bar{\sigma}_n \to \bar{\sigma},\quad \bar{\psi}_n \to \bar{\psi},\quad \mathcal{F}_1[\bar{p}_n + \bar{q}_n, \bar{w}_n] \to \mathcal{F}_1[\bar{p}+\bar{q},\bar{w}],\\ &\text{all strongly in } L^2(\Omega; C[0,\hat{\tau}]),\end{aligned} \tag{3.70}$$

where $\bar{\sigma}_n$, $\bar{\psi}_n$ have obvious meaning. Also,

$$g(\cdot,\cdot,\bar{\theta}_n) \to g(\cdot,\cdot,\bar{\theta}) \quad \text{strongly in } L^\infty(\Omega_{\hat{\tau}}), \tag{3.71}$$

$$\bar{\sigma}_n \bar{p}_{n,xx} \to \bar{\sigma}\bar{p}_{xx} \quad \text{weakly in } L^2(\Omega_{\hat{\tau}}). \tag{3.72}$$

In addition, owing to (3.18), (3.22), and (3.43), the sequence $\{\mathcal{F}_1[\bar{p}_n + \bar{q}_n, \bar{w}_n]_t\}$ is bounded in $L^\infty(0,\hat{\tau}; L^2(\Omega))$, so that we may assume that

$$\mathcal{F}_1[\bar{p}_n + \bar{q}_n, \bar{w}_n]_t \to y \quad \text{weakly-star in } L^\infty(0,\hat{\tau}; L^2(\Omega)) \tag{3.73}$$

for some $y \in L^\infty(0,\hat{\tau}; L^2(\Omega))$. But then it follows from (3.70) that $y = \mathcal{F}_1[\bar{p}+\bar{q},\bar{w}]_t$. Finally, we conclude from (3.67) and (3.68) that

$$\bar{p}^2_{n,xx} \to \bar{p}^2_{xx} \quad \text{weakly in } L^2(\Omega_{\hat{\tau}}). \tag{3.74}$$

Indeed, we have for any test function $\eta \in C_0^\infty(\Omega_{\hat{\tau}})$ that

$$\begin{aligned}&\lim_{n\to\infty} \int_0^{\hat{\tau}} \int_\Omega \bar{p}^2_{n,xx}\eta \,dx\,dt\\ &= -\lim_{n\to\infty} \int_0^{\hat{\tau}} \int_\Omega \Big(\bar{p}_{n,xxx}\bar{p}_{n,x}\eta + \bar{p}_{n,xx}\bar{p}_{n,x}\eta_x\Big)\,dx\,dt\\ &= -\int_0^{\hat{\tau}} \int_\Omega \Big(\bar{p}_{xxx}\bar{p}_x\eta + \bar{p}_{xx}\bar{p}_x\eta_x\Big)\,dx\,dt = \int_0^{\hat{\tau}} \int_\Omega \bar{p}^2_{xx}\eta\,dx\,dt.\end{aligned} \tag{3.75}$$

Since $C_0^\infty(\Omega_{\hat{\tau}})$ is a dense subset of $L^2(\Omega_{\hat{\tau}})$, and since $\{\bar{p}^2_{n,xx}\}$ is bounded in $L^2(\Omega_{\hat{\tau}})$ (cf. (3.47)), we conclude (3.74) from the Banach–Steinhaus theorem.

Now observe that (3.66) implies, in particular, the convergences

$$\theta_{n,t} \to \hat{\theta}_t,\ \ \theta_{n,xx} \to \hat{\theta}_{xx},\ \ p_{n,t} \to \hat{p}_t,\ \ p_{n,xx} \to \hat{p}_{xx},\ \ q_{n,t} \to \hat{q}_t,\ \ w_{n,t} \to \hat{w}_t, \tag{3.76}$$

all weakly in $L^2(\Omega_{\hat{\tau}})$. Combining all the above convergences, and letting $n \to \infty$, we finally can infer that $(\hat{p},\hat{q},\hat{\theta},\hat{w}) = \mathcal{T}(\bar{p},\bar{q},\bar{\theta},\bar{w})$. This concludes the proof of the lemma. □

By virtue of Lemmas 3.3 and 3.4, we deduce from Lemma 3.2 that $\mathcal{T}$ has a fixed point in $\mathcal{M}_{\hat\tau}$ which then is a solution to the system (3.2)–(3.10). To conclude the proof of Theorem 3.1, we still need to show the uniqueness. We achieve this through the following result, which even shows global uniqueness.

LEMMA 3.5. *Let the assumptions of Theorem* 3.1 *be fulfilled, and let* $\tau \in (0,T]$ *be arbitrary. Then the system* (3.2)–(3.10) *has at most one solution in* X_τ.

Proof. Suppose that $(p_i, q_i, \theta_i, w_i) \in X_\tau$, $i = 1, 2$, are two solutions to (3.2)–(3.10) on Ω_τ for some $\tau \in (0,T]$. Let $p := p_1 - p_2$, $q := q_1 - q_2$, $\theta := \theta_1 - \theta_2$, $w := w_1 - w_2$, and put $\sigma_i := \mathcal{H}_1[p_i + q_i, w_i] + \theta_i\, \mathcal{H}_2[p_i + q_i, w_i]$, $\psi_i := \mathcal{H}_3[p_i + q_i, w_i] + \theta_i\, \mathcal{H}_4[p_i + q_i, w_i]$ for $i = 1, 2$. Then it holds that

$$p_t - p_{xx} = \sigma_1 - \sigma_2, \tag{3.77}$$

$$q_t = \sigma_2 - \sigma_1, \tag{3.78}$$

$$\begin{aligned} \theta_t - \theta_{xx} = {} & -\mathcal{F}_1[p_1 + q_1, w_1]_t + \mathcal{F}_1[p_2 + q_2, w_2]_t + p_{1,xx}^2 - p_{2,xx}^2 \\ & + \sigma_1 p_{1,xx} - \sigma_2\, p_{2,xx} + g(x,t,\theta_1) - g(x,t,\theta_2), \end{aligned} \tag{3.79}$$

$$w_t = \psi_1 - \psi_2, \tag{3.80}$$

with corresponding zero initial and boundary conditions. Owing to (H4)(iii), we have for every $(x,t) \in \overline{\Omega_\tau}$

$$\begin{aligned} & \max\{|\sigma_1(x,t) - \sigma_2(x,t)|, |\psi_1(x,t) - \psi_2(x,t)|\} \\ & \le C_1\Big(|\theta(x,t)| + \max_{0\le r\le t}(|p(x,r)| + |q(x,r)| + |w(x,r)|)\Big) \\ & \le C_1\left(|\theta(x,t)| + \int_0^t \Big(|p_t(x,r)| + |q_t(x,r)| + |w_t(x,r)|\Big)\, dr\right), \end{aligned} \tag{3.81}$$

where by C_i, $i \in \mathbb{N}$, we denote positive constants that depend only on the data of the system and on the X_τ-norms of $(p_i, q_i, \theta_i, w_i)$, $i = 1, 2$. Hence, we may multiply (3.77) by p_t, and by $-p_{xx}$, respectively, (3.78) by q_t, and (3.80) by w_t, respectively, add the four resulting equations, integrate over space and time, and apply Young's inequality appropriately to arrive at the estimate

$$\begin{aligned} \int_0^t \int_\Omega \Big(p_t^2 + p_{xx}^2 + q_t^2 + w_t^2\Big)\, dx\, ds \le {} & C_2 \int_0^t \int_0^s \int_\Omega \Big(p_t^2 + q_t^2 + w_t^2\Big)\, dx\, dr\, ds \\ & + C_3 \int_0^t \int_\Omega \theta^2\, dx\, ds. \end{aligned} \tag{3.82}$$

Next, we integrate (3.79) over $[0,s]$ for some $s > 0$. We obtain

$$\begin{aligned} \theta - \int_0^s \theta_{xx}\, dr = {} & -\mathcal{F}_1[p_1 + q_1, w_1] + \mathcal{F}_1[p_2 + q_2, w_2] + \int_0^s \Big(p_{1,xx}^2 - p_{2,xx}^2\Big)\, dr \\ & + \int_0^s \Big(\sigma_1 p_{1,xx} - \sigma_2 p_{2,xx}\Big)\, dr + \int_0^s \Big(g(x,r,\theta_1) - g(x,r,\theta_2)\Big)\, dr. \end{aligned} \tag{3.83}$$

Our aim is to multiply (3.83) by θ, and integrate over $\Omega \times [0,t]$ for some $t \in [0,\tau]$. In what follows, we first estimate the terms resulting from the right-hand side individually. To this end, let $\gamma > 0$ (to be specified later). First, we note that for a.e. $(x,t) \in \Omega_\tau$ it holds that

$$|\mathcal{F}_1[p_1+q_1,w_1](x,t)-\mathcal{F}_1[p_2+q_2,w_2](x,t)| \le C_4 \int_0^t \Big(|p_t|+|q_t|+|w_t|\Big)(x,r)\,dr, \tag{3.84}$$

so that, using Young's inequality,

$$\begin{aligned} &\int_0^t \int_\Omega |\theta(x,s)|\,|\mathcal{F}_1[p_1+q_1,w_1](x,s)-\mathcal{F}_1[p_2+q_2,w_2](x,s)|\,dx\,ds \\ &\le \frac{\gamma}{2}\int_0^t\int_\Omega \theta^2\,dx\,ds + \frac{C_5}{2\gamma}\int_0^t\int_0^s\int_\Omega \Big(p_t^2+q_t^2+w_t^2\Big)\,dx\,dr\,ds. \end{aligned} \tag{3.85}$$

Moreover, owing to (H3), we have

$$\begin{aligned} &\int_0^t \int_\Omega |\theta(x,t)| \int_0^s |g(x,r,\theta_1(x,r))-g(x,r,\theta_2(x,r)|\,dr\,dx\,ds \\ &\le C_6 \int_0^t \int_\Omega |\theta(x,t)| \int_0^s |\theta(x,r)|\,dr\,dx\,ds \\ &\le \frac{\gamma}{2}\int_0^t\int_\Omega \theta^2\,dx\,ds + \frac{C_6}{2\gamma}\int_0^t\int_0^s\int_\Omega \theta^2\,dx\,dr\,ds. \end{aligned} \tag{3.86}$$

Next, we estimate

$$\begin{aligned} &\int_0^t \int_\Omega |\theta(x,s)| \int_0^s |\sigma_1 p_{1,xx}-\sigma_2 p_{2,xx}|(x,r)\,dr\,dx\,ds \\ &\le \int_0^t \int_\Omega |\theta(x,s)| \int_0^s |\sigma_2(x,r)|\,|p_{xx}(x,r)|\,dr\,dx\,ds \\ &\quad + \int_0^t \int_\Omega |\theta(x,s)| \int_0^s \Big(|p_{1,xx}|\,|\sigma_1-\sigma_2|\Big)(x,r)\,dr\,dx\,ds \\ &=: B_1 + B_2. \end{aligned} \tag{3.87}$$

By the boundedness of σ_2,

$$B_1 \le \frac{\gamma}{2}\int_0^t\int_\Omega \theta^2\,dx\,ds + \frac{C_7}{2\,\gamma}\int_0^t\int_0^s\int_\Omega p_{xx}^2\,dx\,dr\,ds. \tag{3.88}$$

Next, we employ the Gagliardo–Nirenberg inequality (2.21) with $p=+\infty$, $q=r=2$, $\omega=1/2$, to conclude that, for $i=1,2$ and every $x \in \bar\Omega$,

$$\begin{aligned} &\int_0^s |p_{i,xx}(x,r)|^2\,dr \le \int_0^s \|p_{i,xx}(\cdot,r)\|_\infty^2\,dr \\ &\le 2K_0^2 \int_0^s \Big(\|p_{i,xx}(\cdot,r)\|_2^2 + \|p_{i,xx}(\cdot,r)\|_2 \|p_{i,xxx}(\cdot,r)\|_2\Big)\,dr \\ &\le C_8 \max_{0\le r\le s} \|p_{i,xx}(\cdot,r)\|_2^2 + C_9 \max_{0\le r\le s} \|p_{i,xx}(\cdot,r)\|_2 \left(\int_0^s\int_\Omega |p_{i,xxx}|^2\,dx\,dr\right)^{1/2} \\ &\le C_{10}. \end{aligned} \tag{3.89}$$

Hence, we have

$$\int_0^s \Big(|\sigma_1-\sigma_2|\,|p_{1,xx}|\Big)(x,r)\,dr \le \bigg(\int_0^s |(\sigma_1-\sigma_2)(x,r)|^2\,dr\bigg)^{1/2}\bigg(\int_0^s |p_{1,xx}(x,r)|^2\,dr\bigg)^{1/2}$$
$$\le \sqrt{C_{10}}\,\bigg(\int_0^s |(\sigma_1-\sigma_2)(x,r)|^2\,dr\bigg)^{1/2}, \tag{3.90}$$

so that, by virtue of (3.81) and of Young's inequality,

$$\begin{aligned} B_2 &\le \frac{\gamma}{2}\int_0^t\int_\Omega \theta^2\,dx\,ds + \frac{C_{11}}{2\gamma}\int_0^t\int_0^s\int_\Omega |\sigma_1-\sigma_2|^2\,dx\,dr\,ds \\ &\le \frac{\gamma}{2}\int_0^t\int_\Omega \theta^2\,dx\,ds + \frac{C_{12}}{2\gamma}\int_0^t\int_0^s\int_\Omega \Big(\theta^2+p_t^2+q_t^2+w_t^2\Big)\,dx\,dr\,ds. \end{aligned} \tag{3.91}$$

Finally, using (3.89), we estimate

$$\begin{aligned} &\int_0^t\int_\Omega |\theta(x,s)|\int_0^s \Big(p_{1,xx}^2-p_{2,xx}^2\Big)(x,r)\,dr\,dx\,ds \\ &\le \int_0^t\int_\Omega |\theta(x,s)|\int_0^s \Big(|p_{xx}|\,|p_{1,xx}+p_{2,xx}|\Big)(x,r)\,dr\,dx\,ds \\ &\le \int_0^t\int_\Omega |\theta(x,s)|\bigg(\int_0^s |p_{xx}(x,r)|^2 dr\bigg)^{1/2}\bigg(\int_0^s |(p_{1,xx}+p_{2,xx})(x,r)|^2 dr\bigg)^{1/2} dx\,ds \\ &\le C_{13}\int_0^t\int_\Omega |\theta(x,s)|\,\bigg(\int_0^s |p_{xx}(x,r)|^2\,dr\bigg)^{1/2}\,dx\,ds \\ &\le \frac{\gamma}{2}\int_0^t\int_\Omega \theta^2\,dx\,ds + \frac{C_{14}}{2\gamma}\int_0^t\int_0^s\int_\Omega p_{xx}^2\,dx\,dr\,ds. \end{aligned} \tag{3.92}$$

Now, we multiply (3.83) by θ and integrate over $\Omega\times[0,t]$ for some $t\in[0,\tau]$. Combining the estimates (3.85)–(3.92) and choosing $\gamma>0$ appropriately small, we obtain the inequality

$$\int_0^t\int_\Omega \theta^2\,dx\,ds \le C_{15}\int_0^t\int_0^s\int_\Omega \Big(\theta^2+p_{xx}^2+p_t^2+q_t^2+w_t^2\Big)\,dx\,dr\,ds. \tag{3.93}$$

Consequently, combining inequalities (3.82) and (3.93), we have finally shown that

$$\begin{aligned} &\int_0^t\int_\Omega \Big(\theta^2+p_t^2+p_{xx}^2+q_t^2+w_t^2\Big)\,dx\,ds \\ &\le C_{16}\int_0^t\int_0^s\int_\Omega \Big(\theta^2+p_t^2+p_{xx}^2+q_t^2+w_t^2\Big)\,dx\,dr\,ds, \end{aligned} \tag{3.94}$$

whence, by Gronwall's lemma, $p_t=q_t=w_t=\theta=0$ a.e. in Ω_τ, so that the assertion follows. With this, the proof of Theorem 3.1 is complete. □

4. Global existence. Suppose now that the hypotheses (H1) to (H5) hold so that (3.2)–(3.11) has a unique solution (p,q,θ,w) on $\Omega_{\hat\tau}$ which satisfies (3.11)–(3.14). Using the compatibility condition $u_0(0)=0$, we then easily verify that (u,θ,w), where

$$u(x,t) = \int_0^x \Big(p(\xi,t)+q(\xi,t)\Big)\,d\xi, \tag{4.1}$$

solves (1.1)–(1.7) on $\Omega_{\hat\tau}$ and satisfies (2.16). Now let $\tau \in (0,T]$ be arbitrary such that (u,θ,w) can be extended to a solution of (1.1)–(1.7) on Ω_τ and satisfies $\theta(x,t) \geq \bar\theta$ for some $\bar\theta > 0$, as well as the smoothness properties (2.16). Owing to the global uniqueness result of Lemma 3.5, this solution is unique. We are now going to derive a number of global a priori estimates. To this end, we denote by C_i, $i \in \mathbb{N}$, positive constants which may depend on the given data of system (1.1)–(1.7), but neither on τ nor on the lower bound $\bar\theta$ for the temperature. For notational convenience, we put $\varepsilon := u_x$.

First estimate. We multiply (1.1) by u_t, add the result to (1.3), and then integrate over Ω_t, $t \in (0,\tau]$, and by parts. In light of (H1), we have

$$\begin{aligned} &\frac12 \int_\Omega u_t^2(x,t)\,dx + \int_\Omega \Big(\theta(x,t) + \mathcal{F}_1[\varepsilon,w](x,t)\Big)\,dx \\ &\leq C_1 + \int_0^t \int_\Omega g(x,r,\theta(x,r))\,dx\,dr + \int_0^t \int_\Omega f u_t\,dx\,dr. \end{aligned} \tag{4.2}$$

Invoking (2.2), (2.4), (H2), (H4)(i), the positivity of θ, and Gronwall's lemma, we find that

$$\max_{0\leq t\leq\tau} \Big(\|\theta(\cdot,t)\|_1 + \|u_t(\cdot,t)\|_2 \Big) \leq C_2. \tag{4.3}$$

Second estimate. We multiply (1.3) by $-\theta^{-1}$ and integrate over Ω_t, $t \in (0,\tau]$. (Note that θ^{-1} is actually bounded, since $\theta \geq \bar\theta > 0$.) It follows that

$$\begin{aligned} \int_0^t \int_\Omega \left(\frac{\theta_x^2}{\theta^2} + \frac{\varepsilon_t^2}{\theta} \right) dx\,dr &\leq C_3 + \int_0^t \int_\Omega \frac{1}{\theta} \Big(\mathcal{F}_1[\varepsilon,w]_t - \sigma\varepsilon_t - g(x,r,\theta) \Big)\,dx\,dr \\ &\quad + \int_\Omega \log(\theta(x,t))\,dx. \end{aligned} \tag{4.4}$$

In light of (4.3) and of the elementary inequality $\log(\theta) \leq \theta$ for $\theta > 0$, the second integral on the right-hand side is bounded. Also, we obtain from (1.2), (1.4), (H3), (H5), and Young's inequality that a.e. in Ω_τ it holds that

$$\begin{aligned} &\mathcal{F}_1[\varepsilon,w]_t - \sigma\varepsilon_t - g(x,r,\theta) \leq \mathcal{H}_3[\varepsilon,w]\,\mathcal{G}[w]_t - \theta\,\mathcal{H}_2[\varepsilon,w]\,\varepsilon_t - g_0(x,r) + K_1\theta \\ &\leq -(\theta\,\mathcal{H}_4[\varepsilon,w] + w_t)\,\mathcal{G}[w]_t + K_2\theta|\varepsilon_t| + K_1\theta \leq K_1\theta + K_2\theta|\varepsilon_t| + \frac{K_5}{4} K_2^2\theta^2. \end{aligned} \tag{4.5}$$

Therefore, using (4.3), we find from Young's inequality that

$$\begin{aligned} \int_0^t \int_\Omega \frac{1}{\theta} \Big(\mathcal{F}_1[\varepsilon,w]_t - \sigma\varepsilon_t - g(x,r,\theta) \Big)\,dx\,dr &\leq C_4 \Big(1 + \int_0^t \int_\Omega |\varepsilon_t|\,dx\,dr \Big) \\ &\leq C_5 + \frac12 \int_0^t \int_\Omega \frac{\varepsilon_t^2}{\theta}\,dx\,dr. \end{aligned} \tag{4.6}$$

In conclusion, we have shown the estimate

$$\int_0^\tau \int_\Omega \left(\frac{\theta_x^2}{\theta^2} + \frac{\varepsilon_t^2}{\theta} \right) dx\,dt \leq C_6. \tag{4.7}$$

But then, using Schwarz's inequality for a fixed $t \in (0,\tau)$,

$$\begin{aligned}\int_\Omega \left|\Big(\sqrt{\theta}\Big)_x\right|(x,t)\,dx &= \int_\Omega \left|\frac{\theta_x(x,t)}{2\sqrt{\theta(x,t)}}\right| dx \\ &\le \frac{1}{2}\left(\int_\Omega \theta(x,t)\,dx\right)^{1/2}\left(\int_\Omega \frac{\theta_x^2(x,t)}{\theta^2(x,t)}\,dx\right)^{1/2},\end{aligned} \tag{4.8}$$

whence, using the Gagliardo–Nirenberg inequality (2.21) with $w = \sqrt{\theta}$, $p = +\infty$, $q = 2$, $r = 1$, and $\omega = 1$, and invoking (4.3), we obtain

$$\int_0^\tau \|\theta(\cdot,t)\|_\infty\,dt \le C_7 + C_8\int_0^\tau\int_\Omega \frac{\theta_x^2}{\theta^2}\,dx\,dt \le C_9. \tag{4.9}$$

Hence,

$$\int_0^\tau\int_\Omega \theta^2\,dx\,dt \le \max_{0\le t\le\tau}\|\theta(\cdot,t)\|_1\int_0^\tau \|\theta(\cdot,t)\|_\infty\,dt \le C_{10}. \tag{4.10}$$

Third estimate. We now exploit the decomposition (3.1). We have $u_x = \varepsilon = p + q$, where, owing to (4.3), $\|p\|_{L^\infty(\Omega_\tau)} \le C_{11}$. Therefore, invoking (H4) and (2.11), it holds that for every $(x,t) \in \overline{\Omega_\tau}$,

$$|\sigma(x,t)| + |\psi(x,t)| \le C_{12} + 2K_2\,\theta(x,t) + 2K_4\max_{0\le r\le t}\Big(|q(x,r)| + |w(x,r)|\Big). \tag{4.11}$$

Now multiply (3.5) by q, and (3.8) by w, respectively, add the resulting equations, and integrate over $[0,t]$, where $t \in (0,\tau]$. Using Young's inequality and invoking (4.9), we obtain from (4.11) the estimate

$$\begin{aligned}&\frac{1}{2}\Big(q^2(x,t) + w^2(x,t)\Big)\\ &\le C_{13}\left[1 + \max_{0\le r\le t}\Big(|q(x,r)| + |w(x,r)|\Big)\left(1 + \int_0^t\Big(|q(x,r)| + |w(x,r)|\Big)\,dr\right)\right]\\ &\le C_{14} + \frac{1}{4}\max_{0\le r\le t}\Big(q^2(x,r) + w^2(x,r)\Big) + C_{15}\int_0^t\Big(q^2(x,r) + w^2(x,r)\Big)\,dr.\end{aligned} \tag{4.12}$$

Taking the maximum with respect to t on both sides, we obtain from Gronwall's lemma that

$$\|q\|_{L^\infty(\Omega_\tau)} + \|w\|_{L^\infty(\Omega_\tau)} \le C_{16}, \tag{4.13}$$

whence also

$$\|\varepsilon\|_{L^\infty(\Omega_\tau)} + \max_{j\in\{1,3\}}\|\mathcal{H}_j[\varepsilon,w]\|_{L^\infty(\Omega_\tau)} \le C_{17}, \tag{4.14}$$

and we obtain from (3.5) and (3.8), using (4.9)–(4.11), that

$$\begin{aligned}&\|q_t\|_{L^1(0,\tau;L^\infty(\Omega))\cap L^2(\Omega_\tau)\cap L^\infty(0,\tau;L^1(\Omega))}\\ &\qquad + \|w_t\|_{L^1(0,\tau;L^\infty(\Omega))\cap L^2(\Omega_\tau)\cap L^\infty(0,\tau;L^1(\Omega))} \le C_{18}.\end{aligned} \tag{4.15}$$

Moreover, (4.10) and (4.14) imply that the right-hand side of (3.2) is bounded in $L^2(\Omega_\tau)$; hence, using standard parabolic estimates, we can conclude that

$$\|p\|_{H^1(0,\tau;L^2(\Omega))\cap L^2(0,\tau;H^2(\Omega))\cap C([0,\tau];H^1(\Omega))} \le C_{19}, \tag{4.16}$$

which yields, in particular, that $u_{xt} = p_{xx}$ is bounded in $L^2(\Omega_\tau)$.

Fourth estimate. In the next step, we perform a classical estimate (cf. [5]); namely, we multiply (1.3) by $-\theta^{-1/3}$ and integrate over Ω_t for $t \in (0,\tau]$. It then follows from (4.5) that

$$\int_0^t \int_\Omega \left(\theta^{-4/3}\theta_x^2 + \theta^{-1/3}\varepsilon_t^2\right) dx\, dr \le C_{20}\left(1 + \int_\Omega \theta^{2/3}(x,t)\, dx + \int_0^t \int_\Omega \theta^{2/3}\, dx\, dr\right)$$
$$+ C_{21}\left(\int_0^t \int_\Omega \theta^{2/3}|\varepsilon_t|\, dx\, dr + \int_0^t \int_\Omega \theta^{5/3}\, dx\, dr\right). \tag{4.17}$$

Owing to (4.3) and (4.10), the first, second, and fourth integrals on the right of (4.17) are bounded, and the remaining expression is estimated as follows:

$$\begin{aligned}\int_0^t \int_\Omega \theta^{2/3}|\varepsilon_t|\, dx\, dr &= \int_0^t \int_\Omega \theta^{5/6}\theta^{-1/6}|\varepsilon_t|\, dx\, dr \\ &\le \frac{1}{2}\int_0^t \int_\Omega \theta^{-1/3}\varepsilon_t^2\, dx\, dr + \frac{1}{2}\int_0^t \int_\Omega \theta^{5/3}\, dx\, dr.\end{aligned} \tag{4.18}$$

Since the second summand on the right of (4.18) is again bounded, we have shown the estimate

$$\int_0^\tau \int_\Omega \left(\theta^{-4/3}\theta_x^2 + \theta^{-1/3}\varepsilon_t^2\right) dx\, dt \le C_{22}. \tag{4.19}$$

But then $\theta^{1/3}$ is bounded in $L^\infty(0,\tau;L^3(\Omega)) \cap L^2(0,\tau;H^1(\Omega))$, and the Gagliardo–Nirenberg inequality (2.21), with $p = +\infty$, $r = 2$, $q = 3$, and $\omega = 2/5$, yields that

$$\int_0^\tau \|\theta(\cdot,t)\|_\infty^{5/3}\, dt \le C_{23}, \tag{4.20}$$

whence, using (4.3) once more, we obtain

$$\int_0^\tau \int_\Omega \theta^{8/3}\, dx\, dt \le C_{24}. \tag{4.21}$$

Thus, the right-hand sides of (3.2), (3.5), and (3.8), respectively, are bounded in $L^{8/3}(\Omega_\tau)$, and we can infer from standard parabolic estimates, using $\varepsilon_t = p_{xx}$, that

$$\|p_t\|_{L^{8/3}(\Omega_\tau)} + \|\varepsilon_t\|_{L^{8/3}(\Omega_\tau)} + \|q_t\|_{L^{8/3}(\Omega_\tau)} + \|w_t\|_{L^{8/3}(\Omega_\tau)} \le C_{25}. \tag{4.22}$$

Fifth estimate. We now turn our attention to (1.3). By virtue of (4.21) and (4.22), and invoking (2.5), we easily verify that $\mathcal{F}_1[\varepsilon,w]_t$ and the right-hand side of (1.3) are bounded in $L^{4/3}(\Omega_\tau)$. Therefore, multiplying (1.3) by θ, integrating over Ω_τ for $t \in (0,\tau]$, and applying Young's inequality and (4.3), we see that for any $\gamma > 0$ it holds that

$$\begin{aligned}\|\theta(\cdot,t)\|_2^2 + \int_0^t \int_\Omega \theta_x^2\, dx\, dr &\le C_{26}\left(1 + \gamma^{-1}\right) + \gamma \int_0^t \int_\Omega \theta^4\, dx\, dr \\ &\le C_{27}\left(1 + \gamma^{-1} + \gamma \int_0^t \|\theta(\cdot,r)\|_\infty^3\, dr\right).\end{aligned} \tag{4.23}$$

Now we use (4.3) and the Gagliardo–Nirenberg inequality (2.21) with $p = +\infty$, $q = 1$, $r = 2$, and $\omega = 2/3$ in order to obtain that

$$\int_0^t \|\theta(\cdot,r)\|_\infty^3\, dr \le C_{28}\left(1 + \int_0^t \int_\Omega \theta_x^2\, dx\, dr\right). \tag{4.24}$$

Hence, choosing $\gamma > 0$ small enough, we have shown the estimate

$$\|\theta\|_{L^\infty(0,\tau;L^2(\Omega))\cap L^2(0,\tau;H^1(\Omega))} \le C_{29}, \tag{4.25}$$

whence, using interpolation once more, we obtain

$$\|\theta\|_{L^6(\Omega_\tau)} \le C_{30}. \tag{4.26}$$

Thus, just as in the derivation of (4.22), we have

$$\|p_t\|_{L^6(\Omega_\tau)} + \|\varepsilon_t\|_{L^6(\Omega_\tau)} + \|q_t\|_{L^6(\Omega_\tau)} + \|w_t\|_{L^6(\Omega_\tau)} \le C_{31}. \tag{4.27}$$

But then $\mathcal{F}_1[\varepsilon, w]_t$ and the right-hand side of (1.3) are bounded in $L^2(\Omega_\tau)$, and standard parabolic estimates, also using the continuous imbeddings $H^1(0,\tau;L^2(\Omega)) \cap L^2(0,\tau;H^2(\Omega)) \hookrightarrow C([0,\tau];H^1(\Omega)) \hookrightarrow C(\overline{\Omega_\tau})$, yield that

$$\|\theta\|_{H^1(0,\tau;L^2(\Omega))\cap L^2(0,\tau;H^2(\Omega))\cap C([0,\tau];H^1(\Omega))\cap C(\overline{\Omega_\tau})} \le C_{32}. \tag{4.28}$$

This implies, in particular, that σ_t and ψ_t are bounded in $L^2(\Omega_\tau)$, so that

$$\|q_{tt}\|_{L^2(\Omega_\tau)} + \|w_{tt}\|_{L^2(\Omega_\tau)} \le C_{33}. \tag{4.29}$$

In addition, we may argue as in the derivation of (3.38) to conclude that also

$$\|p\|_{H^2(0,\tau;V_0^*)\cap H^1(0,\tau;H^1(\Omega))} \le C_{34}. \tag{4.30}$$

Sixth estimate. In light of the above estimates, we have, for a.e. $(x,t) \in \Omega_\tau$,

$$\begin{aligned} |\sigma_x(x,t)| + |\psi_x(x,t)| &\le C_{35}\Big(|\theta_x(x,t)| + \max_{0\le r\le t}((|p_x| + |q_x| + |w_x|)(x,r))\Big) \\ &\le C_{36}\left(1 + |\theta_x(x,t)| + \int_0^t \Big(|p_{xt}(x,r)| + |q_{xt}(x,r)| + |w_{xt}(x,r)|\Big)\,dr\right). \end{aligned} \tag{4.31}$$

Hence, differentiating (3.5) and (3.8), respectively, with respect to x, and invoking the estimates (4.28) and (4.30), we easily derive the estimate

$$\|q_{xt}\|_{L^2(\Omega_t)} + \|w_{xt}\|_{L^2(\Omega_t)} \le C_{37}\left(1 + \int_0^t \Big(\|q_{xt}\|_{L^2(\Omega_r)} + \|w_{xt}\|_{L^2(\Omega_r)}\Big)\,dr\right), \tag{4.32}$$

whence, using Gronwall's lemma, we have

$$\|q_{xt}\|_{L^2(\Omega_\tau)} + \|w_{xt}\|_{L^2(\Omega_\tau)} \le C_{38}. \tag{4.33}$$

Finally, we conclude from the above estimates that also

$$\|p_{xxx}\|_{L^2(\Omega_\tau)} \le C_{39}. \tag{4.34}$$

In conclusion, combining all previously shown estimates, we have shown that

$$\|(p,q,\theta,w)\|_{X_\tau} \le C_{40}, \tag{4.35}$$

where X_τ is the space introduced in (3.15).

Conclusion of the proof of Theorem 2.1. So far we have shown that (4.35) holds as long as there is some $\bar\theta > 0$ such that $\theta \ge \bar\theta$ on $\overline{\Omega_\tau}$. To conclude the proof of the assertion, we still have to prove the validity of (2.17). To this end, first observe that we have shown above that $\varepsilon_t = p_{xx}$ is bounded in $L^6(\Omega_\tau) \cap L^2(0,\tau;H^1(\Omega))$. In particular, there is some $\beta_1 > 0$, independent of τ, such that

$$\int_0^t \|\varepsilon_t(\cdot,r)\|_\infty \, dr \le \beta_1 \quad \forall t \in (0,\tau]. \tag{4.36}$$

Besides, there is some $\beta_2 > 0$, independent of τ, such that

$$\max_{(x,t)\in\overline{\Omega_\tau}} |w_t(x,t)| \le \beta_2. \tag{4.37}$$

Now test (1.3) with an arbitrary function $z \in H^1(\Omega_\tau)$ satisfying $z \le 0$ a.e. in Ω_τ. In view of (2.10), (4.5), and (4.37) it follows, for a.e. $t \in (0,T)$, that

$$\begin{aligned}
&\int_\Omega \Big(z\theta_t + z_x\theta_x\Big)(x,t)\,dx \le \int_\Omega |z(x,t)|\Big((\mathcal{F}[\varepsilon,w]_t - \sigma\varepsilon_t - g(\cdot,\cdot,\theta))(x,t)\Big)\,dx \\
&\le \int_\Omega |z(x,t)|\Big((-\theta\,\mathcal{H}_4[\varepsilon,w] - w_t)\,\mathcal{G}[w]_t + K_2\theta|\varepsilon_t| + K_1\theta\Big)(x,t)\,dx \\
&\le \Big(K_1 + K_2K_5\beta_2 + K_2\|\varepsilon_t(\cdot,t)\|_\infty\Big)\int_\Omega |z(x,t)|\,\theta(x,t)\,dx \\
&\le \varphi(t)\int_\Omega |z(x,t)|\,\theta(x,t)\,dx,
\end{aligned} \tag{4.38}$$

where $\varphi(t) := (K_1 + K_2K_5\beta_2 + K_2\|\varepsilon_t(\cdot,t)\|_\infty)$ is by (4.36) bounded in $L^1(0,\tau)$ by a constant which does not depend on $\tau \in (0,T]$. Now, put

$$z(x,t) := -\left(\delta\exp\left(-\int_0^t \varphi(s)\,ds\right) - \theta(x,t)\right)^+ \quad \text{for } (x,t)\in\Omega_\tau. \tag{4.39}$$

Then it follows from inequality (4.38) that

$$\begin{aligned}
&\int_\Omega \left(z\left(z + \delta\exp\left(-\int_0^t \varphi(s)\,ds\right)\right)_t + z_x^2\right)(x,t)\,dx \\
&\le \varphi(t)\int_\Omega |z|\left(|z| + \delta\exp\left(-\int_0^t \varphi(s)\,ds\right)\right)(x,t)\,dx.
\end{aligned} \tag{4.40}$$

This yields, in particular,

$$\frac{1}{2}\,\frac{d}{dt}\int_\Omega z^2(x,t)\,dx + \int_\Omega z_x^2(x,t)\,dx \le \varphi(t)\int_\Omega z^2(x,t)\,dx. \tag{4.41}$$

Therefore, by Gronwall's lemma, $z = 0$, and thus

$$\theta(x,t) \ge \delta e^{-((K_1+K_2K_5\beta_2)t + K_2\beta_1)} \quad \forall (x,t)\in\overline{\Omega_\tau}. \tag{4.42}$$

Therefore, we can claim that $\tau = T$, and the assertion of Theorem 2.1 is completely proved. □

Remark 4. It does not present any major difficulties to extend the above proof to the more general case when $\mathcal{H}_3$ and $\mathcal{H}_4$ are *vector* hysteresis operators and, accordingly, (1.4) is a vector differential equation (then, of course, the hypotheses (H4) and (H5) have to be appropriately modified).

Remark 5. It is easy to see that the solution (u, θ, w) depends Lipschitz continuously on the data of the system. Indeed, a closer look at the proof of Lemma 3.5 reveals that $L^2(\Omega)$-variations of u_0, u_1, θ_0, w_0 and $L^2(\Omega_T)$-variations of f lead to Lipschitz variations of (p, q, θ, w) in the norm of the space $(H^1(0,T;L^2(\Omega)) \cap L^2(0,T;H^2(\Omega))) \times H^1(0,T;L^2(\Omega)) \times L^2(\Omega_T) \times H^1(0,T;L^2(\Omega))$. A similar result holds for variations of g. As the line of arguments should be clear, we leave the explicit formulation and the proof of the corresponding result to the reader.

Remark 6. It seems natural to discuss the asymptotic behavior of system (1.1)–(1.7) as $\mu \searrow 0$. As our method of proof strongly relies on the presence of the viscous term $-\mu u_{xxt}$ in (1.1), our analysis does not cover this problem, which seems to be very difficult.

REFERENCES

[1] G. Andrews, *On the existence of solutions to the equation $u_{tt} = u_{xxt} + \sigma(u_x)_x$*, J. Differential Equations, 35 (1980), pp. 200–231.

[2] M. Brokate and J. Sprekels, *Hysteresis and Phase Transitions*, Appl. Math. Sci. 121, Springer-Verlag, New York, 1996.

[3] G. Caginalp, *An analysis of a phase field model of a free boundary*, Arch. Ration. Mech. Anal., 92 (1986), pp. 205–245.

[4] C. M. Dafermos, *Global smooth solutions to the initial-boundary value problem for the equations of one-dimensional nonlinear thermoviscoelasticity*, SIAM J. Math. Anal., 13 (1982), pp. 397–408.

[5] C. Dafermos and L. Hsiao, *Global smooth thermomechanical processes in one-dimensional thermoviscoelasticity*, Nonlinear Anal., 6 (1982), pp. 435–454.

[6] R. E. Edwards, *Functional Analysis*, Holt, Rinehart and Winston, New York, 1965.

[7] G. Gilardi, P. Krejčí, and J. Sprekels, *A hysteresis approach to phase-field models with thermal memory*, Math. Methods Appl. Sci., 23 (2000), pp. 909–922.

[8] M. A. Krasnosel'skii and A. V. Pokrovskii, *Systems with Hysteresis*, Springer-Verlag, Heidelberg, 1989 (Russian edition: Nauka, Moscow, 1983).

[9] P. Krejčí, *Hysteresis, Convexity and Dissipation in Hyperbolic Equations*, GAKUTO Internat. Ser. Math. Sci. Appl. 8, Gakkōtosho, Tokyo, 1996.

[10] P. Krejčí and J. Sprekels, *On a system of nonlinear PDEs with temperature-dependent hysteresis in one-dimensional thermoplasticity*, J. Math. Anal. Appl., 209 (1997), pp. 25–46.

[11] P. Krejčí and J. Sprekels, *Temperature-dependent hysteresis in one-dimensional thermo-visco-elastoplasticity*, Appl. Math., 43 (1998), pp. 173–206.

[12] P. Krejčí and J. Sprekels, *Hysteresis operators in phase-field models of Penrose-Fife type*, Appl. Math., 43 (1998), pp. 207–222.

[13] P. Krejčí and J. Sprekels, *A hysteresis approach to phase-field models*, Nonlinear Anal., 39 (2000), pp. 569–586.

[14] P. Krejčí and J. Sprekels, *Phase-field models with hysteresis*, J. Math. Anal. Appl., 252 (2000), pp. 198–219.

[15] P. Krejčí and J. Sprekels, *Phase-field systems and vector hysteresis operators*, in Free Boundary Problems: Theory and Applications II, GAKUTO Internat. Ser. Math. Sci. Appl. 14, N. Kenmochi, ed., Gakkōtosho, Tokyo, 2000, pp. 295–310.

[16] P. Krejčí and J. Sprekels, *Phase-field systems for multidimensional Prandtl-Ishlinskii operators with non-polyhedral characteristics*, Math. Methods Appl. Sci., 25 (2002), pp. 309–325.

[17] P. Krejčí, J. Sprekels, and S. Zheng, *Asymptotic behaviour for a phase-field system with hysteresis*, J. Differential Equations, 175 (2001), pp. 88–107.

[18] P. Krejčí, J. Sprekels, and U. Stefanelli, *One-dimensional thermoviscoplastic processes with hysteresis and phase transitions*, Math. Methods Appl. Sci., submitted.

[19] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems*, Vol. I, Springer-Verlag, Heidelberg, 1972.

[20] I. D. Mayergoyz, *Mathematical Models for Hysteresis*, Springer-Verlag, New York, 1991.

[21] Verein Deutscher Eisenhüttenleute, ed., *Steel. A Handbook for Materials Research and Engineering. Vol.* 1: *Fundamentals*, Springer-Verlag, Berlin, 1992.

[22] A. Visintin, *Differential Models of Hysteresis*, Appl. Math. Sci. 111, Springer-Verlag, New York, 1994.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 435–459

ON THE INITIAL VALUE PROBLEM FOR THE ONE DIMENSIONAL QUASI-LINEAR SCHRÖDINGER EQUATIONS*

WEE KEONG LIM† AND GUSTAVO PONCE†

Abstract. We study the local in time solvability of the initial value problem (IVP) of the one dimensional fully nonlinear Schrödinger equation. Under appropriate assumptions on the nonlinearity (regularity and ellipticity) and on the initial data (regularity and decay at infinity), we establish the existence and uniqueness of solutions of the IVP in weighted Sobolev spaces. The equation can be reduced to its quasi-linear version by taking space derivative. The desired results are obtained by combining a change of variables, energy estimates, and the artificial viscosity method.

Key words. quasi-linear Schrödinger equations, a priori estimates

AMS subject classifications. Primary, 35Q55, 35B45; Secondary, 35A07, 35J10

PII. S0036141001399520

1. Introduction.

This paper is concerned with the initial value problem (IVP) for the general quasi-linear Schrödinger equation in one space dimension

$$\begin{cases} \partial_t u = ia(u,\bar{u},\partial_x u,\partial_x\bar{u})\partial_x^2 u + ib(u,\bar{u},\partial_x u,\partial_x\bar{u})\partial_x^2\bar{u} \\ \qquad\qquad + c(u,\bar{u},\partial_x u,\partial_x\bar{u})\partial_x u + d(u,\bar{u},\partial_x u,\partial_x\bar{u})\partial_x\bar{u} + f(u,\bar{u}), \\ u(x,0) = u_0(x), \end{cases} \tag{1.1}$$

where $u = u(x,t)$, $(x,t) \in \mathbb{R}^2$, $a, b, c, d : \mathbb{C}^4 \to \mathbb{C}$, and $f : \mathbb{C}^2 \to \mathbb{C}$ are smooth functions with $a(\cdot)$ a real-valued function. We shall assume the following ellipticity condition:

$$\begin{cases} \text{for any } R > 0, \text{ there exists } \lambda = \lambda(R) > 0 \text{ such that if} \\ \|(z_1,z_2,z_3,z_4)\| = \left(\sum_{j=1}^{4} |z_j|^2\right)^{\frac{1}{2}} \le R, \\ \text{then} \quad \pm a(z_1,z_2,z_3,z_4) - |b(z_1,z_2,z_3,z_4)| \ge \lambda. \end{cases} \tag{1.2}$$

Our goal is to establish a local theory (including existence and uniqueness) in the classical Sobolev spaces $H^s(\mathbb{R})$ or its weighted version $H^s(\mathbb{R}) \cap L^2(|x|^r dx)$, depending on the degree of nonlinearity of a, b, c, d.

Equations of this kind arise in several fields in physics (see [1], [2], and references therein) and have received considerable attention in recent publications.

In [2], de Bouard, Hayashi, and Saut proved local wellposedness for the IVP associated with the equation

$$\partial_t u = i\Delta u - 2iuh'(|u|^2)\Delta(h(|u|^2)) + iug(|u|^2) \tag{1.3}$$

in space dimensions $n = 1, 2, 3$ corresponding to small data u_0 in $H^6(\mathbb{R})$. They also deduced in dimensions $n = 2, 3$ sufficient conditions of the data u_0 which guarantee that the local solution extends to a global one.

*Received by the editors November 22, 2001; accepted for publication (in revised form) April 16, 2002; published electronically December 3, 2002.
http://www.siam.org/journals/sima/34-2/39952.html

†Department of Mathematics, University of California, Santa Barbara, CA 93106 (wklim@math.ucsb.edu, ponce@math.ucsb.edu). The first author's research was conducted while on leave from Multimedia University, Department of Engineering, Jalan Multimedia, 63100 Cyberjaya, Selangor, Malaysia. The second author was supported in part by an NSF grant.

In [18], Poppenberg studied the IVP for the equation of the form

$$\partial_t u = i(\Delta - V(x))u + iuh'(|u|^2)\Delta(h(|u|^2)) + iug(|u|^2). \tag{1.4}$$

Under certain conditions on the potential V and on h and g, he showed that the corresponding IVP is locally wellposed in $H^\infty(\mathbb{R}^n)$ for any dimension n. His proof is based on the Nash–Moser implicit function theorem.

In [5] and [6], Colin removed the smallness condition on the data assumed in [2] to establish the local wellposedness of the IVP for (1.4) in any dimension in $H^s(\mathbb{R}^n)$, $s \geq s_0(n)$.

For the one dimensional case, Poppenberg [19] showed that the IVP associated with the fully nonlinear Schrödinger equation

$$\begin{cases} i\partial_t u = F(t, x, u, \bar{u}, u_x, \bar{u}_x, u_{xx}, \bar{u}_{xx}), \\ u(x, 0) = u_0(x) \end{cases} \tag{1.5}$$

is locally wellposed in $H^\infty(\mathbb{R})$, under appropriate assumptions on F. As in [18], the proof in [19] is based on the Nash–Moser implicit function theorem, which allows it to overcome "the loss of derivatives" introduced by the nonlinearity.

The assumptions in [19] include ellipticity and a "cubic" character of the nonlinearity. For example, the assumptions in [19] exclude equations of the form

$$(a)\ F = u_{xx} + uu_x \ \text{ or } \ (b)\ F = (1 + |u|^2)u_{xx} + i\mathrm{Re}(u)u_x. \tag{1.6}$$

The case (a) above was considered by Ozawa [17]. The implicit "cubic" assumption on the nonlinearity in [19] can be explained by the Mizohata condition [16] which we now describe.

It was shown in [16] that for the local wellposedness of

$$\begin{cases} \partial_t u = i\Delta u + \vec{b}(x) \cdot \nabla u, \quad x \in \mathbb{R}^n, t \in \mathbb{R}, \\ u(x, 0) = u_0(x) \in L^2(\mathbb{R}^n) \end{cases} \tag{1.7}$$

the following condition is necessary:

$$\sup_{x \in \mathbb{R}^n, \omega \in \mathbb{S}^{n-1}} \left| \int_0^\infty Im\ \vec{b}(x + s\omega) \cdot \omega ds \right| < \infty. \tag{1.8}$$

In the one dimensional case, the condition (1.8) applied to the equations in (1.6) (a) suggests that $u(\cdot, t) \in L^1(\mathbb{R})$ for $t \in [0, T]$. However, this does not follow directly from the initial data since even the group $\{e^{it\partial_x^2} : t \in \mathbb{R}\}$ does not preserve the L^1-class.

Notice that when the "nonlinearity" is "cubic," e.g., $F = u_{xx} + u^2 u_x$, the condition (1.8) is "fulfilled" if $u(\cdot, t) \in L^2(\mathbb{R})$.

To state our result, we first need some definition. Let μ_j, $j = 1, 2, 3, 4$, be the smallest integer such that

$$\begin{cases} \partial_z^\alpha a(0,0,0,0) \neq 0, & 0 < |\alpha| \leq \mu_1, \\ \partial_z^\beta b(0,0,0,0) \neq 0, & 0 < |\beta| \leq \mu_2 \ \text{ if } \ b \not\equiv 0, \\ \partial_z^\gamma c(0,0,0,0) \neq 0,\ c(0,0,0,0) = 0, & 0 < |\gamma| \leq \mu_3 \ \text{ if } \ c \not\equiv 0, \\ \partial_z^\sigma d(0,0,0,0) \neq 0,\ d(0,0,0,0) = 0, & 0 < |\sigma| \leq \mu_4 \ \text{ if } \ d \not\equiv 0, \\ f(0,0) = 0, \end{cases} \tag{1.9}$$

where $a = a(z_1, z_2, z_3, z_4)$, similarly for b, c, d and $z = (z_1, z_2, z_3, z_4)$. Let

$$\mu = \min\{\mu_1, \mu_2, \mu_3, \mu_4\}. \tag{1.10}$$

Thus the case $\mu = 1$ corresponds to the "quadratic" case in (1.1) and $\mu \geq 2$ to at least the cubic case in (1.1). Consider the following examples:

Example 1. $a = 1 + z_1^2 + z_2^2 + z_3^2 z_4^2$, $b = z_1 z_2$, $c = z_1 + z_2^2$, $d = z_1 + z_1 z_2 z_3$; then $\mu_1 = \mu_2 = 2$, $\mu_3 = \mu_4 = 1$, so $\mu = 1$, which is a quadratic case.

Example 2. $a = 1 + \text{Re } u + (\text{Re } u)^2$, $b = c = d = 0$ is "quadratic," i.e., $\mu = 1$.

Example 3. $a = 1 + |u_x|^2$, $b = c = d = 0$ is "cubic," i.e., $\mu = 2$.

THEOREM 1. *Assuming* (1.2) *and* (1.9) *with* $\mu \geq 2$, *there exists* $k \in \mathbb{Z}^+$ *such that for any* $u_0 \in H^k(\mathbb{R})$ *there exist* $T = T(\|u_0\|_{H^k})$ *and a unique solution* $u = u(x,t)$ *of the IVP* (1.1) *on the time interval* $[-T, T]$ *such that*

$$u \in C([-T,T] : H^{k-1}(\mathbb{R})) \cap C^1([-T,T] : H^{k-3}(\mathbb{R})).$$

THEOREM 2. *Assuming* (1.2) *and* (1.9) *with* $\mu = 1$, *there exist* $k, r \in \mathbb{Z}^+$ *such that the condition of Theorem* 1 *holds in the space* $H^{k-1}(\mathbb{R}) \cap L^2(|x|^r dx)$ *instead of* $H^{k-1}(\mathbb{R})$.

Remarks.

1. It will follow from our proof that under appropriate assumptions the same results apply to the IVP for the fully nonlinear Schrödinger equation

$$\partial_t w = iF(w, \bar{w}, w_x, \bar{w}_x, w_{xx}, \bar{w}_{xx}). \tag{1.11}$$

   One has to observe that just by taking derivative in (1.11) and using the notation $u = \partial_x w$, one gets an equation similar to (1.1).
   The same applies to the more general form

$$F = F(x, t, w, \bar{w}, w_x, \bar{w}_x, w_{xx}, \bar{w}_{xx}). \tag{1.12}$$

2. As mentioned above, Theorem 2 deals with quadratic nonlinearity for which one "needs" the weighted Sobolev spaces to handle the integrability conditions in (1.8).
3. Our ellipticity assumption (1.2) allows us to consider equations of the form $a(\cdot) = \frac{1}{1+|u|^2} = 1 - \frac{|u|^2}{1+|u|^2}$, $b = c = d = f = 0$ which are not "uniformly elliptic."
4. Once the results in Theorem 1 and 2 are established, one can obtain the persistence property by combining these results with those found in [15]. The solution u in Theorems 1 and 2 satisfies $u \in C([-T,T] : H^k(\mathbb{R}))$, and some kind of stability (continuous dependence of the solution upon the initial data) can be established. However, to simplify the exposition, we shall not pursue these results here.
5. We do not attempt to get the optimal value of the parameter s (in H^s) and r (in $L^2(|x|^r dx)$) in Theorems 1 and 2 provided by our arguments below.

Our method of proof consists of several steps.

First we differentiate j times the equation in (1.1) to obtain an equation for $v_j = \partial_x^j u = \frac{\partial^j u}{\partial x^j}$, $j = 0, 1, \ldots, k-1$, whose coefficients for the second order derivatives are similar to those in (1.1).

We perform energy estimates in these equations which depend on one higher derivative, i.e., $v_{j+1} = \partial_x^{j+1} u$. To close the estimate we need to bound $v_k = \partial_x^k u$. To

do this, we shall use the "gauge transformation" introduced by Hayashi and Ozawa in [11]. However, this process presents the difficulty of working with $\phi\partial_x^k u$ instead of $\partial_x^k u$, where ϕ is the factor arising from the "gauge transformation" which has exponential form and depends only on some derivatives $\partial_x^j u$, $j = 0, 1, 2, 3$.

To overcome this difficulty, we used ϕ as part of our unknown and show that both norms for $\phi\partial_x^k u$ and $\partial_x^k u$ are equivalent in some time interval.

Putting all these together we get an a priori estimate of the form

$$\frac{d}{dt}h(t) \leq \left((h(t))^2 + (h(t))^l\right) e^{(h(t))^2+(h(t))^l}, \tag{1.13}$$

at least the cubic case (Theorem 1), with l depending on the highest nonlinearity considered.

From (1.13) it follows that there exists $T = T(\|u_0\|_{H^k}) > 0$ such that, for $t \in [0, T]$, $h(t) \leq 100h(0)$ and in this time interval $h(t) \cong \|u(t)\|_{H^k}$. This provides to the norm an a priori estimate.

To prove existence we rely on the artificial viscosity method, so we need to check that the previous argument is preserved for the one parameter family of parabolic equations whose solutions converge as the parameter defining the parabolic character goes to zero.

At this point we explain the restriction of the dimension. In [14], Kenig, Ponce, and Vega established the local wellposedness of the IVP associated with the equation

$$\partial_t u = i\Delta u + P(u, \bar{u}, \nabla_x u, \nabla_x \bar{u}), \tag{1.14}$$

where $P : \mathbb{C}^{2n+2} \to \mathbb{C}$ is a polynomial without constant or linear terms, under the smallness assumption for the data. In [11], Hayashi and Ozawa used a gauge transformation to remove in one dimension the smallness assumption on the data in [14].

In [3] and [4], using ideas due to Doi [8], Chihara was able to remove the smallness condition in [14] in any dimension. Roughly speaking, in [4], Chihara relied on a symmetrization process based on the ellipticity of the dispersive factor in (1.14) and the gauge transformation, which in the case $n \geq 2$ involves pseudodifferential operators. In the one dimensional case, this symmetrization process is not necessary. Also the fact that the "gauge transformation" is exact, i.e., it does not involve pseudodifferential operator, is crucial in our result. The results of Chihara in [4] have been improved by Hayashi and Kaikina [10].

The proof in [14] is based on the following smoothing effects of the solution of the lineal IVP:

$$\begin{cases} \partial_t u = i\Delta u + f(x,t), \\ u(x,0) = u_0(x). \end{cases}$$

If $f \equiv 0$, then

$$\int_0^T \int_{\mathbb{R}^n} |D_x^{1/2} e^{it\Delta} u_0|^2 \lambda(x) dx dt \leq c\|u_0\|_{L^2}, \tag{1.15}$$

and if $u_0 \equiv 0$, then

$$\int_0^T \int_{\mathbb{R}^n} \left| \partial_x \int_0^t e^{i(t-t')\Delta} f(\cdot, t') dt' \right|^2 \lambda(x) dx dt \leq c \int_0^T \int_{\mathbb{R}^n} |f(x)|^2 \lambda^{-1}(x) dx dt, \tag{1.16}$$

where $\lambda(x)$ is a weight with appropriate decay at ∞. The inequality (1.15) was proved simultaneously in [7], [20], [21], and (1.16) in [13].

We do not need to establish the smoothing effect in our solutions as part of the proof to get the existence of solutions. The fact that the solutions provided by Theorems 1 and 2 satisfy the "smoothing effect" of the kind described in (1.13)–(1.14) (gain $\frac{1}{2}$-derivative with respect to the initial data and 1-derivative with respect to the inhomogeneous part) follow immediately by the results in [15]; see also [8], [9]. The main point is that once the existence of a (smooth enough) solution has been established, the proof of the corresponding smoothing effect reduces to a linear problem treated in [8], [9], and [15].

The rest of this paper is organized as follows. We will prove Theorem 1 in section 2 and Theorem 2 in section 3. The appendix includes some technical details. All the integrations are with respect to x and over the whole real line $\mathbb{R}$ unless otherwise mentioned.

2. Proof of Theorem 1. We shall divide the proof in several steps.

2.1. Step 1: Formal energy estimate for lower derivatives. Consider the general quasi-linear Schrödinger equation

$$\begin{cases} \partial_t u = ia(u,\bar{u},\partial_x u,\partial_x\bar{u})\partial_x^2 u + ib(u,\bar{u},\partial_x u,\partial_x\bar{u})\partial_x^2\bar{u} \\ \qquad\qquad + c(u,\bar{u},\partial_x u,\partial_x\bar{u})\partial_x u + d(u,\bar{u},\partial_x u,\partial_x\bar{u})\partial_x\bar{u} + f(u,\bar{u}), \\ u(x,0) = u_0(x) \in H^k(\mathbb{R}), \end{cases} \tag{2.1}$$

where $u = u(x,t), (x,t) \in \mathbb{R}^2$. We recall our hypotheses:

1. a, b, c, d, f are smooth functions of its arguments.
2. a is a real-valued function.
3. For any $\lambda > 0$, there exist $m_1, M_1 > 0$ such that if $\|(z_1,z_2,z_3,z_4)\| \le \lambda$, then we have
$$m_1 \le a(z_1,z_2,z_3,z_4) - |b(z_1,z_2,z_3,z_4)| \le M_1,$$
thus there exist $m_2, M_2 > 0$ such that $m_2 \le \frac{1}{a \pm |b|} \le M_2$. (It will be clear from the proof below that the case $-a(z_1,z_2,z_3,z_4) - |b(z_1,z_2,z_3,z_4)|$ is similar.)
4. $a(z_1,z_2,z_3,z_4) - a(0,0,0,0) = O(|z_1|^\alpha + |z_2|^\alpha + |z_3|^\alpha + |z_4|^\alpha)$, $\alpha \ge \mu \ge 2$, similarly for $b(z_1,z_2,z_3,z_4) - b(0,0,0,0)$, $c(z_1,z_2,z_3,z_4), d(z_1,z_2,z_3,z_4)$.
5. We consider only the case $t \in [0,T]$. (The case $t < 0$ follows by a similar argument.)

To perform the energy estimate on $\partial_x^j u$, we need to take the jth derivative of (2.1), $j = 0, 1, \ldots, k$. With the notation $\partial_x^j u = v_j$, we have from (4.8) (Appendix 1) that (2.1) is rewritten as

$$\partial_t v_j = ia\partial_x^2 v_j + ib\partial_x^2\bar{v}_j \ + c_j\partial_x v_j + d_j\partial_x\bar{v}_j + f_j, \tag{2.2}$$

where $a'_j = \partial_{z_j} a(z_1,z_2,z_3,z_4)$, similarly for b'_j, c'_j, d'_j, $j = 1,2,3,4$; $a' = \partial_x(a) = \frac{\partial a}{\partial x}$, where $a = a(u(x,t),\bar{u}(x,t),\partial_x u(x,t),\partial_x\bar{u}(x,t))$ is implicitly a function of x,t, similarly for b', c', d' and

$$\begin{cases} c_j = ija' + ia'_3 v_2 + ib'_3\bar{v}_2 + c + c'_3 v_1 + d'_3\bar{v}_1, \\ d_j = ijb' + ia'_4 v_2 + ib'_4\bar{v}_2 + d + c'_4 v_1 + d'_4\bar{v}_1, \\ f_j = f_{aj} + f_{bj} + f_{cj} + f_{dj} + \partial_x^j f, \end{cases} \tag{2.3}$$

where $f_{aj}, f_{bj}, f_{cj}, f_{dj}$ are described in (4.7) and depend only on at most jth derivatives of the unknown u, i.e., on $v_0, \bar{v}_0, \dots, v_j, \bar{v}_j$. Thus,

$$\begin{cases} c_j = c_j(v_0, \bar{v}_0, v_1, \bar{v}_1, v_2, \bar{v}_2), \\ d_j = d_j(v_0, \bar{v}_0, v_1, \bar{v}_1, v_2, \bar{v}_2), \\ f_j = f_j(v_0, \bar{v}_0, v_1, \bar{v}_1, \dots, v_j, \bar{v}_j). \end{cases} \tag{2.4}$$

Notice that c_j and d_j depend on j only as a multiplicative constant.

For $j \le k-1$, we perform the standard energy estimate on (2.2), i.e., multiply the equation by $\bar{v}_j$, integrate with respect to x over $\mathbb{R}$, and take the real part of the result. After integration by parts, we have

$$\begin{aligned} \frac{d}{dt}\int |v_j|^2 dx =\ & i\int a(\bar{v}_j\partial_x^2 v_j - v_j\partial_x^2\bar{v}_j)dx + i\int(\bar{v}_j b\partial_x^2\bar{v}_j - v_j\bar{b}\partial_x^2 v_j)dx \\ & + \int(c_j\bar{v}_j\partial_x v_j + \bar{c}_j v_j\partial_x\bar{v} + d_j\bar{v}_j\partial_x\bar{v}_j + \bar{d}_j v_j\partial_x v_j)dx + \int(v_j f_j + \bar{v}_j\bar{f}_j)dx \\ =\ & -2i\mathrm{Re}\int \bar{v}_j a' v_{j+1}dx - 2i\mathrm{Re}\int b(\bar{v}_{j+1})^2 dx + i\mathrm{Re}\int b''\bar{v}_j^2 dx \\ & + 2\mathrm{Re}\int c_j\bar{v}_j v_{j+1}dx - \mathrm{Re}\int d_j'\bar{v}_j^2 dx + 2\mathrm{Re}\int v_j f_j dx. \end{aligned} \tag{2.5}$$

Thus, we obtain the following estimate:

$$\begin{aligned} \frac{d}{dt}\|v_j\|_{L^2}^2 \le\ & 2\|b\|_{L^\infty}\|v_{j+1}\|_{L^2}^2 + 2(\|a'\|_{L^\infty} + \|c_j\|_{L^\infty})\|v_j\|_{L^2}\|v_{j+1}\|_{L^2} \\ & + (\|b''\|_{L^\infty} + \|d_j'\|_{L^\infty})\|v_j\|_{L^2}^2 + 2\|f_j\|_{L^2}\|v_j\|_{L^2}. \end{aligned} \tag{2.6}$$

We shall use that $a' = a_1'\partial_x u + a_2'\partial_x\bar{u} + a_3'\partial_x^2 u + a_4'\partial_x^2\bar{u}$ can be viewed as a first order polynomial of $u, \bar{u}, \partial_x u, \partial_x\bar{u}, \partial_x^2 u, \partial_x^2\bar{u}$, i.e., $v_0, \bar{v}_0, v_1, \bar{v}_1, v_2, \bar{v}_2$; similarly, c_j, d_j, and b'' can be viewed as first and second order polynomials of $v_0, \bar{v}_0, v_1, \bar{v}_1, v_2, \bar{v}_2$. Observe that the estimate in (2.6) depends on v_{j+1}. However, for $j = k$, $\frac{d}{dt}\int |v_k|^2 dx$ depends on v_{k+1}. To have the estimate close within $H^k(\mathbb{R})$, we need to consider a gauge transformation.

2.2. Step 2: Estimate for $v_k = \partial_x^k u$, the gauge transformation. In this section, we will use a gauge transformation to get the desired estimate for $v_k = \partial_x^k u$. It is performed as follows: consider the kth derivative of the quasi-linear Schrödinger equation (1.1) and its conjugate as a system and rewrite them in a matrix form. The ellipticity (1.2) $a^2 - |b|^2 > 0$ will guarantee the nonsingularity of the matrix $A = \left(\begin{smallmatrix} a & b \\ -\bar{b} & -a \end{smallmatrix}\right)$. The equation is rewritten in the form which after introducing a weight function ϕ we are able to close the estimate within $H^k(\mathbb{R})$ with some equivalent "norm."

From (2.2) with $j = k$ and its conjugate, we get

$$\begin{cases} \partial_t v_k = ia\partial_x^2 v_k + ib\partial_x^2\bar{v}_k + c_k\partial_x v_k + d_k\partial_x\bar{v}_k + f_k, \\ \partial_t\bar{v}_k = -ia\partial_x^2\bar{v}_k - i\bar{b}\partial_x^2 v_k + \bar{c}_k\partial_x\bar{v}_k + \bar{d}_k\partial_x v_k + \bar{f}_k, \end{cases} \tag{2.7}$$

which can be written as the system

$$\partial_t\begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} = i\begin{pmatrix} a & b \\ -\bar{b} & -a \end{pmatrix}\partial_x^2\begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} + \begin{pmatrix} c_k & d_k \\ \bar{d}_k & \bar{c}_k \end{pmatrix}\partial_x\begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} + \begin{pmatrix} f_k \\ \bar{f}_k \end{pmatrix}. \tag{2.8}$$

Let

$$A=\begin{pmatrix} a & b \\ -\bar{b} & -a \end{pmatrix}, \quad \tilde{a}=\tfrac{a}{a^2-|b|^2}, \quad \tilde{b}=\tfrac{b}{a^2-|b|^2}$$
$$\text{so } A^{-1}=\begin{pmatrix} \tilde{a} & \tilde{b} \\ -\bar{\tilde{b}} & -\tilde{a} \end{pmatrix}. \tag{2.9}$$

Notice that ellipticity guarantees that $a^2-|b|^2>0$.

So (2.8) is equivalent to

$$A^{-1}\partial_t \begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} = i\partial_x^2 \begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} + A^{-1}\begin{pmatrix} c_k & d_k \\ \bar{d}_k & \bar{c}_k \end{pmatrix}\partial_x \begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} + A^{-1}\begin{pmatrix} f_k \\ \bar{f}_k \end{pmatrix}. \tag{2.10}$$

Write

$$\begin{cases} A^{-1}\begin{pmatrix} c_k & d_k \\ \bar{d}_k & \bar{c}_k \end{pmatrix} = \begin{pmatrix} \tilde{a}c_k+\tilde{b}\bar{d}_k & \tilde{a}d_k+\tilde{b}\bar{c}_k \\ -\bar{\tilde{b}}c_k-\tilde{a}\bar{d}_k & -\bar{\tilde{b}}d_k-\tilde{a}\bar{c}_k \end{pmatrix} = \begin{pmatrix} \alpha_{11}^k & \alpha_{12}^k \\ \alpha_{21}^k & \alpha_{22}^k \end{pmatrix}, \\ A^{-1}\begin{pmatrix} f_k \\ \bar{f}_k \end{pmatrix} = \begin{pmatrix} \tilde{a}f_k+\tilde{b}\bar{f}_k \\ -\bar{\tilde{b}}f_k-\tilde{a}\bar{f}_k \end{pmatrix} = \begin{pmatrix} F_k \\ -\bar{F}_k \end{pmatrix}. \end{cases} \tag{2.11}$$

Thus (2.10) becomes

$$\begin{pmatrix} \tilde{a} & \tilde{b} \\ -\bar{\tilde{b}} & -\tilde{a} \end{pmatrix}\partial_t \begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} = i\partial_x^2 \begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} + \begin{pmatrix} \alpha_{11}^k & \alpha_{12}^k \\ \alpha_{21}^k & \alpha_{22}^k \end{pmatrix}\partial_x \begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} + \begin{pmatrix} F_k \\ \bar{F}_k \end{pmatrix}. \tag{2.12}$$

Since c_k, d_k depend on k only as a multiplicative constant, then α_{mn}^k, $m,n=1,2$, depend on k also as a multiplicative constant.

We now apply a gauge transformation to (2.12), i.e., multiply a function ϕ (which will be determined later) to the system (2.12) and write the system in the following form:

$$\begin{aligned} \begin{pmatrix} \tilde{a} & \tilde{b} \\ -\bar{\tilde{b}} & -\tilde{a} \end{pmatrix}\left[\partial_t \begin{pmatrix} v_k\phi \\ \bar{v}_k\phi \end{pmatrix} - \partial_t\phi \begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix}\right] = {} & i\partial_x^2 \begin{pmatrix} v_k\phi \\ \bar{v}_k\phi \end{pmatrix} - 2i\partial_x\phi\partial_x \begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} \\ & - i\partial_x^2\phi \begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} + \begin{pmatrix} \alpha_{11}^k & \alpha_{12}^k \\ \alpha_{21}^k & \alpha_{22}^k \end{pmatrix}\phi\partial_x \begin{pmatrix} v_k \\ \bar{v}_k \end{pmatrix} + \phi \begin{pmatrix} F_k \\ \bar{F}_k \end{pmatrix}. \end{aligned} \tag{2.13}$$

Consider the first equation of (2.13),

$$\begin{aligned} \tilde{a}\partial_t(v_k\phi)+\tilde{b}\partial_t(\bar{v}_k\phi)-\partial_t\phi(\tilde{a}v_k+\tilde{b}\bar{v}_k) = {} & i\partial_x^2(v_k\phi)-2i\partial_x\phi\partial_x v_k \\ & - i(\partial_x^2\phi)v_k+(\alpha_{11}^k\partial_x v_k+\alpha_{12}^k\partial_x\bar{v}_k)\phi+\phi F_k, \end{aligned} \tag{2.14}$$

which is equivalent to

$$\begin{aligned} \tilde{a}\partial_t(v_k\phi)+\tilde{b}\partial_t(\bar{v}_k\phi) = {} & i\partial_x^2(v_k\phi)+(-2i\partial_x\phi+\alpha_{11}^k\phi)\partial_x v_k \\ & + \alpha_{12}^k\phi\partial_x\bar{v}_k+\partial_t\phi(\tilde{a}v_k+\tilde{b}\bar{v}_k)-i(\partial_x^2\phi)v_k+\phi F_k. \end{aligned} \tag{2.15}$$

We want to eliminate the term having $\partial_x v_k$ as a factor so we can perform an appropriate energy estimate. Hence we should choose ϕ such that

$$-2i\partial_x\phi+\alpha_{11}^k\phi=0.$$

Thus

$$\begin{aligned}\phi = \phi(x,t) &= \exp\left(\frac{1}{2i}\int_0^x \left(\alpha_{11}^k\right)(x',t)dx'\right)\\ &= \exp\left(\frac{1}{2i}\int_0^x \left(\tilde{a}c_k + \tilde{b}\bar{d}_k\right)(x',t)dx'\right)\\ &= \exp\left\{\frac{1}{2i}\int_0^x \Big(\tilde{a}(ika' + ia_3'v_2 + ib_3'\bar{v}_2 + \ c + c_3'v_1 + d_3'\bar{v}_1)\right.\\ &\qquad \left.+ \tilde{b}(-ia_4'\bar{v}_2 - ik\bar{b}' - i\bar{b}_4'v_2 + \bar{c}_4'\bar{v}_1 + \bar{d} + \bar{d}_4'v_1)\Big)(x',t)dx'\right\}.\end{aligned} \tag{2.16}$$

Observe the following:

1. $\phi = \phi(v_0, \bar{v}_0, v_1, \bar{v}_1, v_2, \bar{v}_2)$, since a, b, c, d are functions of $v_0, \bar{v}_0, v_1, \bar{v}_1,\ v_2, \bar{v}_2$ and the expression in the exponential involves only one derivative.
2. $\phi = \phi_k$ depends on k only as a multiplicative constant.
3. $\phi^{-1} = \frac{1}{\phi}$.
4. If $b \equiv 0$, then $\phi = e^{\frac{1}{2i}\int_0^x \left(\frac{c_k}{a}\right)(x',t)dx'}$.
5. $\phi = e^{\frac{1}{2i}\int_0^x \left(\tilde{a}c_k + \tilde{b}d_k\right)(x',t)dx'} = e^{\rho + i\nu}$, which gives $|\bar{\phi}\phi^{-1}| = |\phi\bar{\phi}^{-1}| = 1$.
6. For $\mu \geq 2$, $\tilde{a}c_k + \tilde{b}d_k$ is at least quadratic, which means
$$|\phi| \leq e^{C\int |\tilde{a}c_k + \tilde{b}d_k|dx} \leq e^{C(\|u\|_{H^3}^2 + \|u\|_{H^3}^l)},$$
where l depends on the order of a, b, c, d.

With this choice of ϕ, (2.15) is reduced to

$$\begin{aligned}\tilde{a}\partial_t(v_k\phi) + \tilde{b}\partial_t(\bar{v}_k\phi) = i\partial_x^2(v_k\phi) + \alpha_{12}^k\phi\partial_x\bar{v}_k + \partial_t\phi(\tilde{a}v_k + \tilde{b}\bar{v}_k)\\ - i(\partial_x^2\phi)v_k + \phi F_k.\end{aligned} \tag{2.17}$$

We rewrite (2.17) in the following form:

$$\begin{aligned}\tilde{a}\partial_t(v_k\phi) + \tilde{b}\partial_t(\bar{v}_k\phi) &= i\partial_x^2(v_k\phi) + \alpha_{12}^k\phi\bar{\phi}^{-1}\partial_x(\bar{v}_k\bar{\phi}) + (\partial_t\phi)\tilde{a}\phi^{-1}(v_k\phi)\\ &\quad + (\partial_t\phi)\tilde{b}\bar{\phi}^{-1}(\bar{v}_k\bar{\phi}) - \alpha_{12}^k\phi\bar{\phi}^{-2}\partial_x\bar{\phi}(\bar{v}_k\bar{\phi})\\ &\quad - i(\partial_x^2\phi)\phi^{-1}(v_k\phi) + \phi F_k.\end{aligned} \tag{2.18}$$

Performing the energy estimate on (2.18), i.e., multiplying (2.18) by $\bar{v}_k\bar{\phi}$, integrating with respect to x over $\mathbb{R}$, and taking the real part, we have

$$\begin{aligned}&\int [2\tilde{a}\partial_t(|v_k\phi|^2) + \tilde{b}(\bar{v}_k\bar{\phi})\partial_t(\bar{v}_k\phi) + \bar{\tilde{b}}(v_k\phi)\partial_t(v_k\bar{\phi})]dx\\ &\quad = \int i\left[(\bar{v}_k\bar{\phi})\partial_x^2(v_k\phi) - (v_k\phi)\partial_x^2(\bar{v}_k\bar{\phi})\right]dx\\ &\quad\quad + \int \left[\alpha_{12}^k\phi\bar{\phi}^{-1}(\bar{v}_k\bar{\phi})\partial_x(\bar{v}_k\bar{\phi}) + \bar{\alpha}_{12}^k\bar{\phi}\phi^{-1}(v_k\phi)\partial_x(v_k\phi)\right]dx\\ &\quad\quad + 2\mathrm{Re}\int \partial_t\phi\left[\tilde{a}\phi^{-1}(\bar{v}_k\bar{\phi})(v_k\phi) + \tilde{b}\bar{\phi}^{-1}(\bar{v}_k\bar{\phi})(\bar{v}_k\bar{\phi})\right]dx\\ &\quad\quad - 2\mathrm{Re}\int \alpha_{12}^k\phi\bar{\phi}^{-2}\partial_x\bar{\phi}(\bar{v}_k\bar{\phi})(\bar{v}_k\bar{\phi})dx\\ &\quad\quad - 2\mathrm{Re}\int \left[i(\partial_x^2\phi)\phi^{-1}(\bar{v}_k\bar{\phi})(v_k\phi) + \phi(\bar{v}_k\bar{\phi})F_k\right]dx.\end{aligned} \tag{2.19}$$

Using that

$$\tilde{b}(\bar{v}_k\bar{\phi})\partial_t(\bar{v}_k\phi) = \tilde{b}\bar{\phi}\phi^{-1}(\bar{v}_k\phi)\partial_t(\bar{v}_k\phi) = \frac{\tilde{b}\bar{\phi}\phi^{-1}}{2}\partial_t\left((\bar{v}_k\phi)^2\right),$$
$$\bar{\tilde{b}}(v_k\phi)\partial_t(v_k\bar{\phi}) = \frac{\bar{\tilde{b}}\phi\bar{\phi}^{-1}}{2}\partial_t\left((v_k\bar{\phi})^2\right),$$

the left-hand side of (2.19) becomes

$$\begin{aligned}
&\int\left[\tilde{a}\partial_t(|v_k\phi|^2) + \tilde{b}(\bar{v}_k\bar{\phi})\partial_t(\bar{v}_k\phi) + \bar{\tilde{b}}(v_k\phi)\partial_t(v_k\bar{\phi})\right]dx\\
&\quad= \int\left[\tilde{a}\partial_t(|v_k\phi|^2) + \frac{\tilde{b}\bar{\phi}\phi^{-1}}{2}\partial_t\left((\bar{v}_k\phi)^2\right) + \frac{\bar{\tilde{b}}\phi\bar{\phi}^{-1}}{2}\partial_t\left((v_k\bar{\phi})^2\right)\right]dx \qquad (2.20)\\
&\quad= \frac{d}{dt}\int\left[\tilde{a}(|v_k\phi|^2) + \mathrm{Re}[\tilde{b}\bar{\phi}\phi^{-1}(\bar{v}_k\phi)^2]\right]dx\\
&\qquad- \int(\partial_t\tilde{a}|v_k\phi|^2 + \mathrm{Re}[\partial_t(\tilde{b}\bar{\phi}\phi^{-1})(\bar{v}_k\phi)^2])dx.
\end{aligned}$$

Combining (2.20) and integration by parts on the right-hand side of (2.19), it follows that

$$\begin{aligned}
\frac{d}{dt}\int\left[\tilde{a}|v_k\phi|^2 + \mathrm{Re}(\tilde{b}\bar{\phi}\phi^{-1}(\bar{v}_k\phi)^2)\right]dx &= -\mathrm{Re}\int\left[\partial_x\left(\alpha_{12}^k\phi\bar{\phi}^{-1}\right)(\bar{v}_k\bar{\phi})^2\right]dx\\
&+ 2\mathrm{Re}\int\partial_t\phi\left[\tilde{a}\phi^{-1}|v_k\phi|^2 + \tilde{b}\bar{\phi}^{-1}(\bar{v}_k\bar{\phi})^2\right]dx\\
&- 2\mathrm{Re}\int\alpha_{12}^k\phi\bar{\phi}^{-2}\partial_x\bar{\phi}(\bar{v}_k\bar{\phi})^2dx \qquad (2.21)\\
&- 2\mathrm{Re}\int\left[i(\partial_x^2\phi)\phi^{-1}|v_k\phi|^2 + \phi(\bar{v}_k\bar{\phi})F_k\right]dx\\
&+ \int\partial_t\tilde{a}|v_k\phi|^2dx + \mathrm{Re}\int\partial_t(\tilde{b}\bar{\phi}\phi^{-1})(\bar{v}_k\phi)^2dx.
\end{aligned}$$

Note that $\partial_t a, \partial_t b, \partial_t\bar{b}$ can be expressed in terms of a, b, c, d, f, their first order derivatives (i.e., a_1', a_2', a_3', a_4', and so on), and their conjugates. Also, observe that $\partial_t a, \partial_t b, \partial_t\bar{b}$ depend only on $v_0, \bar{v}_0, \ldots, v_3, \bar{v}_3$ (Appendix 2). Thus we have the following estimate:

$$\begin{aligned}
\frac{d}{dt}\int\left[\tilde{a}|v_k\phi|^2 + \mathrm{Re}[\tilde{b}\bar{\phi}\phi^{-1}(\bar{v}_k\phi)^2]\right]dx \;&\le 2\Big\{\|\partial_x\left(\alpha_{12}^k\phi\bar{\phi}^{-1}\right)\|_{L^\infty}\\
&+ \|\partial_t\phi\tilde{a}\phi^{-1}\|_{L^\infty} + \|\partial_t\phi\tilde{b}\bar{\phi}^{-1}\|_{L^\infty} + \|\alpha_{12}^k\phi\bar{\phi}^{-2}\partial_x\bar{\phi}\|_{L^\infty}\\
&+ \|(\partial_x^2\phi)\phi^{-1}\|_{L^\infty} + \|\partial_t\left(b\bar{\phi}\phi^{-1}\right)\|_{L^\infty} + \frac{1}{2}\|\partial_t\tilde{a}\|_{L^\infty}\Big\}\|v_k\phi\|_{L^2}^2 \qquad (2.22)\\
&+ 2\int|\phi F_k\bar{v}_k\bar{\phi}|dx\\
&= J_1\|v_k\phi\|_{L^2}^2.
\end{aligned}$$

The expressions for J_1 depend only on $v_j, \bar{v}_j$ for $j \le 5$ and $|\phi|$ which depend only on $v_0, \bar{v}_0, v_1, \bar{v}_1, v_2, \bar{v}_2$ (Appendix 2); this guarantees an appropriate boundedness of J_1 since $|\phi| \le e^{C(\|u\|_{H^3}^2 + \|u\|_{H^3}^l)}$. Thus, with these properties, we consider the expression

$$|||v|||^2 = \sum_{j=0}^{k-1}\|v_j\|_{L^2}^2 + \left(\int\left[\tilde{a}|v_k\phi|^2 + \mathrm{Re}(\tilde{b}\bar{\phi}\phi^{-1}(\bar{v}_k\phi)^2)\right]dx\right). \qquad (2.23)$$

From our hypothesis $m_2 \leq \frac{1}{a\pm|b|} \leq M_2$, as long as $\|(z_1,z_2,z_3,z_4)\| \leq \lambda$ with $\lambda = 100\|u_0\|_{H^k}$, using that $|\bar{\phi}\phi^{-1}| = 1$, we have

$$\int \left[\tilde{a}|v_k\phi|^2 + \mathrm{Re}(\tilde{b}\bar{\phi}\phi^{-1}(\bar{v}_k\phi)^2)\right]dx \geq \int \left[\tilde{a}|v_k\phi|^2 - |\tilde{b}||v_k\phi|^2\right]dx \tag{2.24}$$
$$= \int \frac{a-|b|}{a^2-|b|^2}|v_k\phi|^2dx \geq m_2\int |v_k\phi|^2dx = m_2\|v_k\phi\|_{L^2}^2.$$

Collecting (2.6), (2.22)–(2.24), we conclude that

$$\frac{d}{dt}|||v||| \leq C(|||v|||^2 + |||v|||^l)e^{C(|||v|||^2+|||v|||^l)}, \tag{2.25}$$

which shows that there exists $T = T(\|u_0\|_{H^k}) > 0$ such that

$$|||v(t)||| \leq 100|||v(0)|||,$$

and, within this range,

$$|||\cdot||| \sim \|\cdot\|_{H^k}. \tag{2.26}$$

At this point, we have only established an a priori estimate for the L^2-norm of $v_0, v_1, \ldots, v_k$. We shall obtain the existence of $v_0, v_1, \ldots, v_k$ by introducing two viscosity terms to our IVP (1.1). This will be done in the next section.

2.3. Step 3: Estimate for v_k in the equation with viscosity. To establish the local existence of (1.1) in $H^k(\mathbb{R})$, we shall consider first a parabolic version of (1.1). The energy estimate for this new equation can be obtained using an argument similar to that provided above. So we should consider only the additional terms coming from the added parabolic part.

To show existence, we need to introduce a modified version of (1.1) with two viscosity terms: for $\varepsilon \in (0,1]$,

$$\begin{cases} \partial_t u = -\varepsilon\partial_x^4 u + \varepsilon\delta\partial_x^2 u + ia\partial_x^2 u + ib\partial_x^2\bar{u} + c\partial_x u + d\partial_x\bar{u} + f, \\ u(x,0) = u_0(x), \end{cases} \tag{2.27}$$

where $\delta > 0$ will be determined.

As before, to establish the energy estimate within $H^k(\mathbb{R})$, we need to take the jth derivative of (2.27) to get the same expression as in (2.2) with two extra terms:

$$\partial_t v_j = -\varepsilon\partial_x^4 v_j + \varepsilon\delta\partial_x^2 v_j + ia\partial_x^2 v_j + ib\partial_x^2\bar{v}_j \ + c_j\partial_x v_j + d_j\partial_x\bar{v}_j + f_j. \tag{2.28}$$

For $j \leq k-1$, perform the energy estimate for (2.27): take the jth derivative of (2.27), multiply by $\bar{v}_j$, integrate with respect to x, and take the real part:

$$\begin{aligned} \frac{d}{dt}\int |v_j|^2dx &= -\varepsilon\int(\bar{v}_j\partial_x^4 v_j + v_j\partial_x^4\bar{v}_j)dx + \varepsilon\delta\int(\bar{v}_j\partial_x^2 v_j + v_j\partial_x^2\bar{v}_j)dx \\ &\quad - 2i\mathrm{Re}\int \bar{v}_j a' v_{j+1}dx - 2i\mathrm{Re}\int b(\bar{v}_{j+1})^2dx + i\mathrm{Re}\int b''\bar{v}^2dx \\ &\quad + 2\mathrm{Re}\int c_j\bar{v}_j v_{j+1}dx - \mathrm{Re}\int d_j'\bar{v}_j^2dx + 2\mathrm{Re}\int v_j f_j dx \\ &= -\varepsilon\int(\bar{v}_j\partial_x^4 v_j + v_j\partial_x^4\bar{v}_j)dx + \varepsilon\delta\int(\bar{v}_j\partial_x^2 v_j + v_j\partial_x^2\bar{v}_j)dx + P_j. \end{aligned} \tag{2.29}$$

We have already discussed all the terms in P_j in (2.6), so we just need to consider the first two terms in the right-hand side of (2.29). After integration by parts, (2.29) becomes

$$\partial_t \int |v_j|^2 dx = -2\varepsilon \int |\partial_x^2 v_j|^2 dx - 2\varepsilon\delta \int |\partial_x v_j|^2 dx + P_j. \tag{2.30}$$

Since $-2\varepsilon \int |\partial_x^2 v_j|^2 dx - 2\varepsilon\delta \int |\partial_x v_j|^2 dx \le 0$, the estimate for the jth equations, for $j \le k-1$, can be handled exactly as in (2.5)–(2.6). For $j = k$, again, we need to consider other alternatives: we will rewrite the equation for $j = k$ and its conjugate in matrix form and again use a gauge transformation to get the desired estimate.

Consider (2.28) when $j = k$ and its conjugate,

$$\begin{cases} \partial_t v_k = -\varepsilon\partial_x^4 v_k + \varepsilon\delta\partial_x^2 v_k + ia\partial_x^2 v_k + ib\partial_x^2 \bar v_k \\ \qquad\qquad + c_k \partial_x v_k + d_k \partial_x \bar v_k + f_k, \\ \partial_t \bar v_k = -\varepsilon\partial_x^4 \bar v_k + \varepsilon\delta\partial_x^2 \bar v_k - ia\partial_x^2 \bar v_k - i\bar b\partial_x^2 v_k \\ \qquad\qquad + \bar c_k \partial_x \bar v_k + \bar d_k \partial_x v_k + \bar f_k. \end{cases} \tag{2.31}$$

Rewrite the equations in matrix form by using (2.9)–(2.11):

$$\begin{aligned} \begin{pmatrix} \tilde a & \tilde b \\ -\bar{\tilde b} & -\tilde a \end{pmatrix} \partial_t \begin{pmatrix} v_k \\ \bar v_k \end{pmatrix} &= \begin{pmatrix} \tilde a & \tilde b \\ -\bar{\tilde b} & -\tilde a \end{pmatrix} (-\varepsilon\partial_x^4 + \varepsilon\delta\partial_x^2) \begin{pmatrix} v_k \\ \bar v_k \end{pmatrix} \\ &\quad + i\partial_x^2 \begin{pmatrix} v_k \\ \bar v_k \end{pmatrix} + \begin{pmatrix} \alpha_{11}^k & \alpha_{12}^k \\ \alpha_{21}^k & \alpha_{22}^k \end{pmatrix} \partial_x \begin{pmatrix} v_k \\ \bar v_k \end{pmatrix} + \begin{pmatrix} F_k \\ -\bar F_k \end{pmatrix}. \end{aligned} \tag{2.32}$$

Consider the first equation of (2.32):

$$\begin{aligned} \tilde a\partial_t v_k + \tilde b\partial_t \bar v_k = \tilde a(-\varepsilon\partial_x^4 + \varepsilon\delta\partial_x^2)v_k + \tilde b(-\varepsilon\partial_x^4 + \varepsilon\delta\partial_x^2)\bar v_k \\ + i\partial_x^2 v_k + (\alpha_{11}^k \partial_x v_k + \alpha_{12}^k \partial_x \bar v_k) + F_k. \end{aligned}$$

Using gauge transformation with the same ϕ as in (2.16), we have

$$\begin{aligned} \tilde a\partial_t(v_k\phi) + \tilde b\partial_t(\bar v_k\phi) = \ & \phi\tilde a(-\varepsilon\partial_x^4 + \varepsilon\delta\partial_x^2)v_k + \phi\tilde b(-\varepsilon\partial_x^4 + \varepsilon\delta\partial_x^2)\bar v_k \\ & + i\partial_x^2(v_k\phi) + \alpha_{12}^k \phi\partial_x \bar v_k + \partial_t\phi(\tilde a v_k + \tilde b\bar v_k) - i(\partial_x^2\phi)v_k + \phi F_k. \end{aligned} \tag{2.33}$$

Perform the energy estimate on the last equation (in divergent form): multiplying by $\bar v_k\bar\phi$, integrating with respect to x over $\mathbb{R}$ and taking the real part, we have

$$\begin{aligned} \frac{d}{dt}\int \left[\tilde a|v_k\phi|^2 + \mathrm{Re}(\tilde b\bar\phi\phi^{-1}(\bar v_k\phi)^2)\right] dx &= 2\mathrm{Re}\int (\bar v_k\bar\phi)\tilde a(-\varepsilon\phi\partial_x^4 v_k + \varepsilon\delta\phi\partial_x^2 v_k)dx \\ &+ 2\mathrm{Re}\int (\bar v_k\bar\phi)\tilde b(-\varepsilon\phi\partial_x^4 \bar v_k + \varepsilon\delta\phi\partial_x^2 \bar v_k)dx - \mathrm{Re}\int \left[\partial_x\left(\alpha_{12}^k\phi\bar\phi^{-1}\right)(\bar v_k\bar\phi)^2\right] dx \\ &+ 2\mathrm{Re}\int \partial_t\phi\left[\tilde a\phi^{-1}|v_k\phi|^2 + \tilde b\bar\phi^{-1}(\bar v_k\bar\phi)^2\right] dx - 2\mathrm{Re}\int \alpha_{12}^k\phi\bar\phi^{-2}\partial_x\bar\phi(\bar v_k\bar\phi)^2 dx \\ &- 2\mathrm{Re}\int \left[i(\partial_x^2\phi)\phi^{-1}|v_k\phi|^2 + \phi(\bar v_k\bar\phi)F_k\right] dx + \int \partial_t\tilde a|v_k\phi|^2 dx \\ &+ \mathrm{Re}\int \partial_t\left(\tilde b\bar\phi\phi^{-1}\right)(\bar v_k\phi)^2 dx. \end{aligned} \tag{2.34}$$

Only the terms

$$2\mathrm{Re}\int(\bar{v}_k\bar{\phi})\tilde{a}(-\varepsilon\phi\partial_x^4 v_k+\varepsilon\delta\phi\partial_x^2 v_k)dx+2\mathrm{Re}\int(\bar{v}_k\bar{\phi})\tilde{b}(-\varepsilon\phi\partial_x^4\bar{v}_k+\varepsilon\delta\phi\partial_x^2\bar{v}_k)dx$$

need to be considered since the other terms have been dealt with in (2.6) and (2.22). It can be shown (Appendix 3) that with an appropriate choice of δ, the terms in (2.34) generated by the two viscosity terms can be absorbed into the energy estimate. Thus we have the following proposition.

PROPOSITION 2.1. *There exist $\delta=\delta(\|u_0\|_{H^k})>0$, $C=C(\delta,\|u_0\|_{H^k})$ such that for $\varepsilon\in(0,1]$*

$$\begin{aligned}(2.35)\qquad 2\mathrm{Re}\int(\bar{v}_k\bar{\phi})\tilde{a}(-\varepsilon\phi\partial_x^4 v_k+\varepsilon\delta\phi\partial_x^2 v_k)dx\\ +2\mathrm{Re}\int(\bar{v}_k\bar{\phi})\tilde{b}(-\varepsilon\phi\partial_x^4\bar{v}_k+\varepsilon\delta\phi\partial_x^2\bar{v}_k)dx\le C\|v_k\phi\|_{L^2}^2.\end{aligned}$$

Gathering (2.6), (2.22), and Proposition 2.1, we have the a priori estimate of the norm $|||\cdot|||$ similar to (2.26) for the solution u^ε of (2.27); i.e., there exists $T=T(\|u_0\|_{H^k})>0$ such that if $u^\varepsilon\in C([0,T]:H^k(\mathbb{R}))$ is a solution of (2.27), then $\|u^\varepsilon(t)\|_{H^k}\le 100\|u_0\|_{H^k}$ for $t\in[0,T]$.

2.4. Step 4: The local existence of a solution with viscosity terms. With the a priori estimate discussed above, we are ready to show the local existence and uniqueness of the IVP (2.27). Consider

$$\begin{aligned}(2.36)\qquad \partial_t u&=-\varepsilon\partial_x^4 u+\varepsilon\delta\partial_x^2 u+ia\partial_x^2 u+ib\partial_x^2\bar{u}+c\partial_x u+d\partial_x\bar{u}+f\\ &=(-\varepsilon\partial_x^4+\varepsilon\delta\partial_x^2)u+F(u,\bar{u},\partial_x u,\partial_x\bar{u},\partial_x^2 u,\partial_x^2\bar{u}).\end{aligned}$$

2.4.1. The linear equation. We will first look at some properties of the operator of the linear homogeneous equation

$$(2.37)\qquad\begin{cases}\partial_t w=-\varepsilon\partial_x^4 w+\varepsilon\delta\partial_x^2 w, & x\in\mathbb{R},t>0,\\ w(0,x)=w_0(x),\end{cases}$$

where $w=w(t,x)$. The IVP (2.37) has the solution

$$(2.38)\qquad \hat{w}(t,\xi)=(e^{(-\varepsilon(2\pi\xi)^4+\varepsilon\delta(2\pi\xi)^2)t})\hat{w}_0(\xi)=(e^{-\varepsilon(2\pi\xi)^4 t}e^{\varepsilon\delta(2\pi\xi)^2 t})\hat{w}_0(\xi).$$

If $\widehat{K}_{\varepsilon\delta t}=e^{\varepsilon\delta(2\pi\xi)^2 t}$ and $\widehat{G}_{\varepsilon t}(\xi)=e^{-\varepsilon(2\pi\xi)^4 t}$, then the solution of (5.2) can be written as

$$(2.39)\qquad w(x,t)=(G_{\varepsilon t}*(K_{\varepsilon\delta t}*w_0))(x).$$

Note that

$$(2.40)\qquad G_{\varepsilon t}(x)=\int_{\mathbb{R}}e^{2\pi i x\xi}e^{-\varepsilon t(2\pi)^4\xi^4}d\xi,$$

so the change of variable $\eta=\xi(\varepsilon t)^{1/4}$ gives

$$\begin{aligned}(2.41)\qquad G_{\varepsilon t}(x)&=\int_{\mathbb{R}}e^{2\pi i x\eta/(\varepsilon t)^{1/4}}e^{-(2\pi\eta)^4}\frac{d\eta}{(\varepsilon t)^{1/4}}\\ &=\frac{1}{(\varepsilon t)^{1/4}}(e^{-(2\pi\eta)^4})^\vee\left(\frac{x}{(\varepsilon t)^{1/4}}\right)\end{aligned}$$

and

$$\partial_x^\alpha G_{\varepsilon t}(x) = \int_{\mathbb{R}} \frac{(2\pi\eta)^\alpha}{(\varepsilon t)^{\alpha/4}} e^{2\pi i x\eta/(\varepsilon t)^{1/4}} e^{-(2\pi\eta)^4} \frac{d\eta}{(\varepsilon t)^{1/4}}. \tag{2.42}$$

We introduce the notation

$$\mathcal{W}(t)f = (e^{-\varepsilon(2\pi\xi)^4 t}\widehat{f}(\xi))^\vee(x). \tag{2.43}$$

2.4.2. The existence of a solution of (2.27). We will prove the existence and uniqueness of a solution to the IVP

$$\begin{cases} \partial_t u = (-\varepsilon\partial_x^4 + \varepsilon\delta\partial_x^2)u + F(u, \bar{u}, \partial_x u, \partial_x\bar{u}, \partial_x^2 u, \partial_x^2\bar{u}), \\ u(x,0) = u_0(x), \end{cases} \tag{2.44}$$

as the fixed point for the operator

$$\begin{aligned} \Phi_\varepsilon(u(t)) &= \mathcal{W}(t)(K_{\varepsilon\delta t} * u_0) + \int_0^t e^{(-\varepsilon\partial_x^4 + \varepsilon\delta\partial_x^2)(t-t')} F(x,t')dt' \\ &= \mathcal{W}(t)(K_{\varepsilon\delta t} * u_0) + \int_0^t \mathcal{W}(t-t')(K_{\varepsilon\delta t} * F)(x,t')dt'. \end{aligned} \tag{2.45}$$

We observe that

$$\begin{aligned} \partial_x^s \Phi_\varepsilon(u(t)) &= \mathcal{W}(t)(K_{\varepsilon\delta t} * \partial_x^s u_0) \\ &\quad + \int_0^t \partial_x^2 \mathcal{W}(t-t')(K_{\varepsilon\delta t} * \partial_x^{s-2} F)(x,t')dt', \end{aligned} \tag{2.46}$$

so from (2.42)–(2.43) one has the following estimates:

$$\begin{aligned} \sup_{[0,T]} \|\Phi_\varepsilon(u(t))\|_{H^k} &\le C\|w_0\|_{H^k} + C\int_0^t (\varepsilon(t-t'))^{-\frac{1}{2}} \|F\|_{H^{k-2}} dt' \\ &\le C\|w_0\|_{H^k} + C\varepsilon^{-\frac{1}{2}} T^{\frac{1}{2}} \sup_{[0,T]} \|F(t)\|_{H^{k-2}} \\ &\le C\|w_0\|_{H^k} + C\varepsilon^{-\frac{1}{2}} T^{\frac{1}{2}} \sup_{[0,T]} \left(\|u(t)\|_{H^k}^2 + \|u(t)\|_{H^k}^l\right) \end{aligned} \tag{2.47}$$

and

$$\begin{aligned} \sup_{[0,T]} \|(\Phi_\varepsilon(u) - \Phi_\varepsilon(v))(t)\|_{H^s} &\le \varepsilon^{-\frac{1}{2}} T^{\frac{1}{2}} \sup_{[0,T]} \|(F(u) - F(v))(t)\|_{H^{s-2}} \\ &\le \varepsilon^{-\frac{1}{2}} T^{\frac{1}{2}} \sup_{[0,T]} \Big(\|u(t)\|_{H^k} + \|v(t)\|_{H^k} + \|u(t)\|_{H^k}^{l-1} \\ &\qquad + \|v(t)\|_{H^k}^{l-1} \Big) \sup_{[0,T]} \|(u-v)(t)\|_{H^{k-2}}. \end{aligned} \tag{2.48}$$

Collecting the above information we have the following proposition.

PROPOSITION 2.2. *There exists $T_\varepsilon = T_\varepsilon(\varepsilon, \|u_0\|_{H^k})$ such that*

$$\Phi : C([0,T_\varepsilon] : H^k(\mathbb{R})) \to C([0,T_\varepsilon] : H^k(\mathbb{R}))$$

is a contraction mapping.

Our next goal is to remove the dependence on ε of the time interval of existence. We have that for each $\varepsilon > 0$ there exists $T_\varepsilon = T_\varepsilon(\varepsilon, \|u_0\|_{H^k}) = O(\varepsilon)$ such that the unique solution $u^\varepsilon(x,t)$ of

$$\begin{cases} \partial_t u = -\varepsilon\partial_x^4 u + \varepsilon\delta\partial_x^2 u + ia\partial_x^2 u + ib\partial_x^2\bar{u} + c\partial_x u + d\partial_x\bar{u} + f, \\ u(x,0) = u_0(x) \end{cases}$$

satisfies $u^\varepsilon \in C([0,T_\varepsilon] : H^k(\mathbb{R}))$.

Using the a priori estimate obtained in section 2.3, we can reapply the local existence argument given above to extend the local solution u^ε to $C([0,T] : H^k(\mathbb{R}))$ for every $\varepsilon > 0$ with T independent of ε. Therefore, we have shown that for every $u_0 \in H^k(\mathbb{R})$ there exists a $T = T(\|u_0\|_{H^k}) > 0$ independent of ε and a unique solution u^ε of

$$\begin{cases} \partial_t u = -\varepsilon\partial_x^4 u + \varepsilon\delta\partial_x^2 u + ia\partial_x^2 u + ib\partial_x^2\bar{u} + c\partial_x u + d\partial_x\bar{u} + f, \\ u(x,0) = u_0(x) \end{cases}$$

satisfying $u^\varepsilon \in C([0,T] : H^k)$ with

$$\sup_{[0,T]} \|u^\varepsilon(t)\|_{H^k} \le C(\|u_0\|_{H^k}), \tag{2.49}$$

with $C(\|u_0\|_{H^k})$ independent of ε.

Next we shall establish the convergence of u^ε to u, which will be a solution of (1.1) as $\varepsilon \to 0$.

To establish this convergence, we consider the equation for u^ε and $u^{\varepsilon'}$:

$$\begin{cases} \partial_t u^\varepsilon = -\varepsilon\partial_x^4 u^\varepsilon + \delta\varepsilon\partial_x^2 u^\varepsilon + ia(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon)\partial_x^2 u^\varepsilon \\ \qquad + ib(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon)\partial_x^2\bar{u}^\varepsilon + c(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon)\partial_x u^\varepsilon \\ \qquad + d(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon)\partial_x\bar{u}^\varepsilon + f(u^\varepsilon, \bar{u}^\varepsilon) \\ \partial_t u^{\varepsilon'} = -\varepsilon'\partial_x^4 u^{\varepsilon'} + \delta\varepsilon'\partial_x^2 u^{\varepsilon'} + ia(u^{\varepsilon'}, \bar{u}^{\varepsilon'}, \partial_x u^{\varepsilon'}, \partial_x\bar{u}^{\varepsilon'})\partial_x^2 u^{\varepsilon'} \\ \qquad + ib(u^{\varepsilon'}, \bar{u}^{\varepsilon'}, \partial_x u^{\varepsilon'}, \partial_x\bar{u}^{\varepsilon'})\partial_x^2\bar{u}^{\varepsilon'} + c(u^{\varepsilon'}, \bar{u}^{\varepsilon'}, \partial_x u^{\varepsilon'}, \partial_x\bar{u}^{\varepsilon'})\partial_x u^{\varepsilon'} \\ \qquad + d(u^{\varepsilon'}, \bar{u}^{\varepsilon'}, \partial_x u^{\varepsilon'}, \partial_x\bar{u}^{\varepsilon'})\partial_x\bar{u}^{\varepsilon'} + f(u^{\varepsilon'}, \bar{u}^{\varepsilon'}) \end{cases} \tag{2.50}$$

and consider that for the difference of $u^\varepsilon, u^{\varepsilon'}, w = w^{\varepsilon,\varepsilon'} = u^\varepsilon - u^{\varepsilon'}$,

$$\begin{aligned} \partial_t w =& -\varepsilon\partial_x^4 w + \varepsilon\delta\partial_x^2 w - (\varepsilon - \varepsilon')\partial_x^4 u^\varepsilon + (\varepsilon - \varepsilon')\delta\partial_x^2 u^\varepsilon \\ &+ ia(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon)\partial_x^2 w + ib(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon)\partial_x^2\bar{w} \\ &+ i\big(a(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon) - a(u^{\varepsilon'}, \bar{u}^{\varepsilon'}, \partial_x u^{\varepsilon'}, \partial_x\bar{u}^{\varepsilon'})\big)\partial_x^2 u^{\varepsilon'} \\ &+ i\big(b(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon) - b(u^{\varepsilon'}, \bar{u}^{\varepsilon'}, \partial_x u^{\varepsilon'}, \partial_x\bar{u}^{\varepsilon'})\big)\partial_x^2\bar{u}^{\varepsilon'} \\ &+ c(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon)\partial_x w + d(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon)\partial_x\bar{w} \\ &+ \big(c(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon) - c(u^{\varepsilon'}, \bar{u}^{\varepsilon'}, \partial_x u^{\varepsilon'}, \partial_x\bar{u}^{\varepsilon'})\big)\partial_x u^{\varepsilon'} \\ &+ \big(d(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon) - d(u^{\varepsilon'}, \bar{u}^{\varepsilon'}, \partial_x u^{\varepsilon'}, \partial_x\bar{u}^{\varepsilon'})\big)\partial_x\bar{u}^{\varepsilon'} \\ &+ f(u^\varepsilon, \bar{u}^\varepsilon) - f(u^{\varepsilon'}, \bar{u}^{\varepsilon'}). \end{aligned} \tag{2.51}$$

By expressing the factors

$$a^\varepsilon - a^{\varepsilon'} = a(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x\bar{u}^\varepsilon) - a(u^{\varepsilon'}, \bar{u}^{\varepsilon'}, \partial_x u^{\varepsilon'}, \partial_x\bar{u}^{\varepsilon'}), \tag{2.52}$$

similarly for $b^\varepsilon - b^{\varepsilon'}, c^\varepsilon - c^{\varepsilon'}, d^\varepsilon - d^{\varepsilon'}$ in an appropriate manner, we can rewrite (2.51) in the following form:

$$\partial_t w = ia^\varepsilon \partial_x^2 w + ib^\varepsilon \partial_x^2 \bar{w} + \tilde{c}\partial_x w + \tilde{d}\partial_x \bar{w} + \tilde{q}_1 w + \tilde{q}_2 \bar{w} + (\varepsilon + \varepsilon')\Psi, \tag{2.53}$$

where $\tilde{c} = \tilde{c}(u^\varepsilon, \bar{u}^\varepsilon, \partial_x u^\varepsilon, \partial_x \bar{u}^\varepsilon, \partial_x^2 u^\varepsilon, \partial_x^2 \bar{u}^\varepsilon, u^{\varepsilon'}, \bar{u}^{\varepsilon'}, \partial_x u^{\varepsilon'}, \partial_x \bar{u}^{\varepsilon'}, \partial_x^2 u^{\varepsilon'}, \partial_x^2 \bar{u}^{\varepsilon'})$, similarly for $\tilde{d}, \tilde{q}_1, \tilde{q}_2, \Psi$. We notice that the last three terms in the right-hand side of (2.53) can be easily handled by the energy estimate. So we will concentrate on the second to the fourth terms. For this we repeat the argument in section 2.2 used to obtain the gauge transformation. The weight function here

$$\phi_{\varepsilon,\varepsilon'} = \phi_{\varepsilon,\varepsilon'}(u^\varepsilon, \ldots, \partial_x^2 \bar{u}^\varepsilon, u^{\varepsilon'}, \ldots, \partial_x^2 \bar{u}^{\varepsilon'}) \tag{2.54}$$

is as in (2.11)–(2.16) such that

$$-2i\partial_x \phi_{\varepsilon,\varepsilon'} + \left(\tilde{a}^\varepsilon \tilde{c} + \tilde{b}^\varepsilon \bar{\tilde{d}}\right)\phi_{\varepsilon,\varepsilon'} = 0, \tag{2.55}$$

where $\tilde{a}^\varepsilon = \frac{a^\varepsilon}{(a^\varepsilon)^2 - |b^\varepsilon|^2}$, $\tilde{b}^\varepsilon = \frac{b^\varepsilon}{(a^\varepsilon)^2 - |b^\varepsilon|^2}$. Thus

$$\phi_{\varepsilon,\varepsilon'} = \exp\left\{\frac{1}{2i}\int_0^x \left(\tilde{a}^\varepsilon \tilde{c} + \tilde{b}^\varepsilon \bar{\tilde{d}}\right)(x', t)dx'\right\}. \tag{2.56}$$

Using that

$$\sup_{[0,T]} \|u^\varepsilon\|_{H^k} \le 100\|u_0\|_{H^k}, \tag{2.57}$$

we know that $\phi_{\varepsilon,\varepsilon'}$ is bounded uniformly on $\varepsilon, \varepsilon' \in (0, 1]$. Combining again the argument in section 2.2 and (2.57), we have

$$\begin{aligned}\frac{d}{dt}\int \left[\tilde{a}^\varepsilon |w\phi_{\varepsilon,\varepsilon'}|^2 + \mathrm{Re}(\tilde{b}^\varepsilon \bar{\phi}_{\varepsilon,\varepsilon'}\phi_{\varepsilon,\varepsilon'}^{-1}(\bar{w}\phi_{\varepsilon,\varepsilon'})^2)\right]dx \\ \le C\|w\phi_{\varepsilon,\varepsilon'}\|_{L^2} + C(\varepsilon - \varepsilon').\end{aligned} \tag{2.58}$$

Using Gronwall inequality, the fact from (2.16), i.e.,

$$\int \left[\tilde{a}^\varepsilon |w\phi_{\varepsilon,\varepsilon'}|^2 + \mathrm{Re}(\tilde{b}^\varepsilon \bar{\phi}_{\varepsilon,\varepsilon'}\phi_{\varepsilon,\varepsilon'}^{-1}(\bar{w}\phi_{\varepsilon,\varepsilon'})^2)\right] dx \ge m_2 \int |w\phi_{\varepsilon,\varepsilon'}|^2 dx, \tag{2.59}$$

where m_2 is independent of $\varepsilon, \varepsilon' \in (0, 1]$ and that $w(0) = 0$, we obtain

$$\lim_{\varepsilon,\varepsilon' \to 0} \sup_{[0,T]} \|w(t)\|_{L^2} = 0. \tag{2.60}$$

Hence $\sup_{[0,T]} \|(u^\varepsilon - u^{\varepsilon'})(t)\|_{L^2} \to 0$ as $\varepsilon, \varepsilon' \to 0$. Consequently, by interpolation, we have

$$\begin{aligned}&\sup_{[0,T]} \|(u^\varepsilon - u^{\varepsilon'})(t)\|_{H^{k-1}} \\ &\qquad \le \sup_{[0,T]} \|(u^\varepsilon - u^{\varepsilon'})(t)\|_{L^2}^{\frac{1}{k}} \sup_{[0,T]} \|(u^\varepsilon - u^{\varepsilon'})(t)\|_{H^k}^{\frac{k-1}{k}}\end{aligned} \tag{2.61}$$

which implies that $\sup_{[0,T]} \|(u^\varepsilon - u^{\varepsilon'})(t)\|_{H^{k-1}} \to 0$; i.e., $\{u^\varepsilon\}_{\varepsilon>0}$ is a Cauchy sequence in the space $C([0,T] : H^{k-1}(\mathbb{R}))$. So there exists $u \in C([0,T] : H^{k-1}(\mathbb{R}))$ which is the limit of the u^ε's as $\varepsilon \to 0$. One also has that $u^\varepsilon(t)$ converges weakly to $u(t)$ in $H^k(\mathbb{R})$ for each $t \in [0,T]$, so $u \in L^\infty([0,T] : H^k(\mathbb{R}))$. Since $k \geq 4$, it is clear that u solves the IVP (1.1).

The proof of the uniqueness in the class described in Theorem 1 follows the same argument given above for the convergence without involving the viscosity part; i.e., if u, v are solutions, we replace u^ε by u and $u^{\varepsilon'}$ by v in (2.51) with $\varepsilon = \varepsilon' = 0$ and repeat the argument.

3. Proof of Theorem 2. The proof of Theorem 2 follows in the same manner as that of Theorem 1; i.e., we establish formal energy estimate for the jth derivative of (1.1) for $j = 1, 2, \ldots, k-1$. For the case $j = k$, we again use a gauge transformation and introduce the same function ϕ in (2.16) to obtain the estimate of the expression (2.24) which is equivalent to $\|\cdot\|_{H^k}$ for some range $|||v(t)||| \leq 100|||v(0)|||$. The problem comes in when the gauge transformation is introduced: ϕ might have some linear terms due to the fact that $a = a(u, \bar{u}, \partial_x u, \partial_x \bar{u})$ or b, c, d may be linear. The weighted $L^2(|x|^r dx)$-norm will allow us to bound $\|\phi\|_{L^\infty}$. We need to consider $H^k \cap L^2(|x|^r dx)$ instead of H^k. Observe that f does not have any influence in the form of ϕ.

In order to proceed, we need the following interpolation lemma.

LEMMA 3.1. *If $u \in H^k(\mathbb{R}) \bigcap L^2(|x|^r dx)$, then $x^\beta \partial_x^\alpha u \in L^2(\mathbb{R})$ with $0 \leq \alpha + \beta \leq r$.*

Next, we observe that the weighted norm $\|\cdot\|_{L^2(|x|^r dx)}$ guarantees the boundedness of $\|\phi\|_{L^\infty}$ for $\mu = 1$; i.e., a, b, c, d might be linear; e.g., let g be some linear term in $ac_k + bd_k$, and then

$$\begin{aligned} \int g(u, \bar{u}, \partial_x u, \partial_x \bar{u}, \partial_x^2 u, \partial_x^2 \bar{u}) dx &= \int \frac{(1+x^2)^{r/2} g}{(1+x^2)^{r/2}} dx \\ &\leq \|(1+x^2)^{r/2} g\|_{L^2} \|(1+x^2)^{-r/2}\|_{L^2} \\ &\leq C \sum_{j=0}^{2} \big(\|\partial_x^j u\|_{L^2(|x|^{2r} dx)} + \|\partial_x^j u\|_{L^2} \big). \end{aligned} \tag{3.1}$$

Hence, for some $l \geq 0$,

$$\begin{aligned} \|\phi\|_{L^\infty} &\leq e^{C \int |ac_k + bd_k| dx} \\ &\leq e^{C \sum_{j \leq 2} \left(\|\partial_x^j u\|_{L^2(|x|^r dx)} + \|\partial_x^j u\|^l_{L^2(|x|^r dx)} \right) + C\left(\|u\|_{H^k} + \|u\|^l_{H^k} \right)}. \end{aligned} \tag{3.2}$$

We will perform the energy estimate for $\|x \partial_x^j u\|_{L^2}$, $j = 0, 1, 2, 3, 4$. We multiply by x the equation in (2.1) to obtain

$$\begin{aligned} \partial_t (xu) = {} & ia \partial_x^2 (xu) - 2ia \partial_x u + ib \partial_x^2 (x\bar{u}) - 2ib \partial_x \bar{u} \\ & + c \partial_x (xu) - cu + d \partial_x (x\bar{u}) - d\bar{u} + xf. \end{aligned} \tag{3.3}$$

To perform energy estimates for xu, we multiply (3.3) by $x\bar{u}$, integrate with respect to x over $\mathbb{R}$ and take the real part:

$$\begin{aligned}\partial_t \int |xu|^2 dx = &-2\mathrm{Re}\int ia' x\bar{u}\partial_x(xu)dx - 4\mathrm{Re}\int iax\bar{u}\partial_x u\ dx\\ &-2\mathrm{Re}\int ib(\partial_x(x\bar{u}))^2 dx - \mathrm{Re}\int ib''(x\bar{u})^2 dx - 4\mathrm{Re}\int ibx\bar{u}\partial_x\bar{u}\ dx\\ &+2\mathrm{Re}\int cx\bar{u}\partial_x(xu)dx - 2\mathrm{Re}\int cx\bar{u}u\ dx\\ &+\mathrm{Re}\int d'(x\bar{u})^2 dx - 2\mathrm{Re}\int dx\bar{u}\bar{u}\ dx - 2\mathrm{Re}\int x\bar{u}xf\ dx\end{aligned} \tag{3.4}$$

which gives

$$\begin{aligned}\partial_t \int |xu|^2 dx \le\ & 2\|a'\|_{L^\infty}\|xu\|_{L^2}\|\partial_x(xu)\|_{L^2} + 4\|a\|_{L^\infty}\|xu\|_{L^2}\|\partial_x u\|_{L^2}\\ &+2\|b\|_{L^\infty}\|\partial_x(xu)\|_{L^2}^2 + \|b''\|_{L^\infty}\|xu\|_{L^2}^2 + 4\|b\|_{L^\infty}\|xu\|_{L^2}\|\partial_x u\|_{L^2}\\ &+2\|c\|_{L^\infty}\|xu\|_{L^2}\|\partial_x(xu)\|_{L^2} + 2\|c\|_{L^\infty}\|xu\|_{L^2}\|u\|_{L^2}\\ &+\|d'\|_{L^\infty}\|xu\|_{L^2}^2 + 2\|d\|_{L^\infty}\|xu\|_{L^2}\|u\|_{L^2} + 2\|xu\|_{L^2}\|xf\|_{L^2}.\end{aligned} \tag{3.5}$$

Thus we can establish the following estimate:

$$\frac{d}{dt}\|xu\|_{L^2} \le C(\|u\|_{H^4} + \|u\|_{H^4}^l)(\|xu\|_{L^2}^2 + \|x\partial_x u\|_{L^2}^2) + C(\|u\|_{H^k}^2 + \|u\|_{H^k}^l) \tag{3.6}$$

for some $l \ge 0$ which depends on the nonlinearity. Similarly, we have

$$\begin{cases}\dfrac{d}{dt}\|x\partial_x u\|_{L^2}^2 \le C(\|u\|_{H^4} + \|u\|_{H^4}^l)(\|xu\|_{L^2}^2 + \|x\partial_x u\|_{L^2}^2 + \|x\partial_x^2 u\|_{L^2}^2)\\ \qquad\qquad + C(\|u\|_{H^k}^2 + \|u\|_{H^k}^l),\\ \dfrac{d}{dt}\|x\partial_x^2 u\|_{L^2}^2 \le C(\|u\|_{H^4} + \|u\|_{H^4}^l)\big(\|xu\|_{L^2}^2 + \|x\partial_x u\|_{L^2}^2\\ \qquad\qquad + \|x\partial_x^2 u\|_{L^2}^2 + \|x\partial_x^3 u\|_{L^2}^2\big) + C(\|u\|_{H^k}^2 + \|u\|_{H^k}^l),\\ \dfrac{d}{dt}\|x\partial_x^3 u\|_{L^2}^2 \le C(\|u\|_{H^4} + \|u\|_{H^4}^l)\big(\|xu\|_{L^2}^2 + \|x\partial_x u\|_{L^2}^2 + \|x\partial_x^2 u\|_{L^2}^2\\ \qquad\qquad + \|x\partial_x^3 u\|_{L^2}^2 + \|x\partial_x^4 u\|_{L^2}^2\big) + C(\|u\|_{H^k}^2 + \|u\|_{H^k}^l),\end{cases} \tag{3.7}$$

which depends on $\|x\partial_x^4 u\|_{L^2}$, i.e., one derivative higher. Thus the estimate for $x\partial_x^4 u$ depends on $x\partial_x^5 u$. We need to consider a gauge transformation to enable the estimate to close within itself which allows us to obtain the desired a priori estimate. We proceed as in section 2.2; consider the equation for $x\partial_x^4 u = xv_4$,

$$\begin{aligned}\partial_t(xv_4) = &\ ia\partial_x^2(xv_4) - 2ia\partial_x v_4 + ib\partial_x^2(x\bar{v}_4) - 2ib\partial_x\bar{v}_4\\ &+ c_4\partial_x(xv_4) - c_4 v_4 + d_4\partial_x(x\bar{v}_4) - d_4\bar{v}_4 + xf_4,\end{aligned} \tag{3.8}$$

where

$$
\begin{aligned}
c_4 &= 4ia' + ia_3'v_2 + ib_3'\bar{v}_2 + c + c_3'v_1 + d_3'\bar{v}_1,\\
d_4 &= 4ib' + ia_4'v_2 + ib_4'\bar{v}_2 + d + c_4'v_1 + d_4'\bar{v}_1,\\
f_4 &= f_{a4} + f_{b4} + f_{c4} + f_{d4} + \partial_x^4 f,\\
f_{a4} &= iv_2\partial_x^3(a_1'v_1 + a_2'\bar{v}_1) + i\sum_{l=2}^{3}\binom{4}{l}(\partial_x^l a)v_{6-l},\\
&\quad + i\sum_{l=1}^{3}\binom{3}{l}v_2\big(\partial_x^l(a_3')v_{5-l} + \partial_x^l(a_4')\bar{v}_{5-l}\big),\\
&\text{similarly for } f_{b4}, f_{c4}, f_{d4} \text{ (Appendix 1)},
\end{aligned}
\tag{3.9}
$$

and its conjugate and write them as the system

$$
\begin{aligned}
\partial_t\begin{pmatrix} xv_4\\ x\bar{v}_4\end{pmatrix} &= i\begin{pmatrix} a & b\\ -\bar{b} & -a\end{pmatrix}\partial_x^2\begin{pmatrix} xv_4\\ x\bar{v}_4\end{pmatrix} + \begin{pmatrix} -2ia + c_4 & -2ib + d_4\\ 2i\bar{d} + \bar{d}_4 & 2ia + \bar{c}_4\end{pmatrix}\partial_x\begin{pmatrix} xv_4\\ x\bar{v}_4\end{pmatrix}\\
&\quad + \begin{pmatrix} -2iav_4 + 2ib\bar{v}_4 - c_4v_4 - d_4\bar{v}_4u + xf_4\\ -2ia\bar{v}_4 - 2i\bar{b}v_4 - \bar{c}_4\bar{v}_4 - \bar{d}_4v_4 + x\bar{f}_4\end{pmatrix}.
\end{aligned}
\tag{3.10}
$$

Using the A as in (2.9) and

$$
A^{-1}\begin{pmatrix} -2ia + c_4 & -2ib + d_4\\ 2i\bar{d} + \bar{d}_4 & 2ia + \bar{c}_4\end{pmatrix} = \begin{pmatrix} \alpha_{11}^4 & \alpha_{12}^4\\ \alpha_{21}^4 & \alpha_{22}^4\end{pmatrix},
\tag{3.11}
$$

the system is written as

$$
A^{-1}\partial_t\begin{pmatrix} xv_4\\ x\bar{v}_4\end{pmatrix} = i\partial_x^2\begin{pmatrix} xv_4\\ x\bar{v}_4\end{pmatrix} + \begin{pmatrix} \alpha_{11}^4 & \alpha_{12}^4\\ \alpha_{21}^4 & \alpha_{22}^4\end{pmatrix}\partial_x\begin{pmatrix} xv_4\\ x\bar{v}_4\end{pmatrix} + A^{-1}\begin{pmatrix} F\\ \bar{F}\end{pmatrix}.
\tag{3.12}
$$

To perform gauge transformation, we multiply a function ϕ_4, which will be determined later to the system, and consider the first equation of the system:

$$
\begin{aligned}
\tilde{a}\partial_t(xv_4\phi_4) + \tilde{b}\partial_t(x\bar{v}_4\phi_4) = {}& i\partial_x^2(xv_4\phi_4) + (\alpha_{11}^4\phi_4 - 2i\partial_x\phi_4)\partial_x(xv_4)\\
&+ \alpha_{12}^4\phi_4\partial_x(x\bar{v}_4) + i(\partial_x^2\phi_4)xv_4 + \phi_4F + \partial_t\phi_4(\tilde{a}xv_4 + \tilde{b}x\bar{v}_4).
\end{aligned}
\tag{3.13}
$$

With the choice of ϕ_4 such that

$$
\begin{cases}\alpha_{11}^4\phi_4 - 2i\partial_x\phi_4 = 0,\\ \phi_4 = \exp\left\{\dfrac{1}{2i}\displaystyle\int_0^x \alpha_{11}^4(x',t)dx'\right\} = e^{\frac{1}{2i}\int_0^x(\tilde{a}(c_4-2ia)+\tilde{b}(d_4+2i\bar{b}))(x',t)dx'},\end{cases}
\tag{3.14}
$$

(3.13) is reduced to

$$
\begin{aligned}
\tilde{a}\partial_t(xv_4\phi_4) + \tilde{b}\partial_t(x\bar{v}_4\phi_4) = {}& i\partial_x^2(xv_4\phi_4) + \alpha_{12}\phi_4\partial_x(x\bar{v}_4) + i(\partial_x^2\phi_4)xv_4\\
&+ \phi_4F + \partial_t\phi_4(\tilde{a}xv_4 + \tilde{b}x\bar{v}_4).
\end{aligned}
\tag{3.15}
$$

Basically ϕ_4 is similar to ϕ in (2.16) except for some extra terms due to the commutator. Also, it has the same properties as ϕ listed after (2.16).

Now we perform the energy estimate: multiply $x\bar{v}_4\bar{\phi}_4$ to (3.15), integrate with respect to x over $\mathbb{R}$ and take the real part. After integration by parts, we have

$$\begin{aligned}\frac{d}{dt}\int \Big[\tilde{a}|xv_4\phi_4|^2 + \mathrm{Re}(\tilde{b}\bar{\phi}\phi_4^{-1}(x\bar{v}_4\phi_4)^2)\Big]dx &= \int(\partial_t\tilde{a})|xu\phi_4|^2dx \\ &+ \int \partial_t(\tilde{b}\bar{\phi}_4\phi_4^{-1})(x\bar{v}_4\phi_4)^2dx - \mathrm{Re}\int \partial_x(\alpha_{12}\bar{\phi}_4\phi_4\bar{\phi}^{-2})(x\bar{v}_4\bar{\phi})^2dx \\ &+ 2\mathrm{Re}\int i(\partial_x^2\phi_4)\phi_4^{-1}|xu\phi_4|^2dx + \int(x\bar{v}_4\bar{\phi}_4)\phi_4F\ dx \\ &+ \int(\partial_t\phi_4)(\tilde{a}\phi_4^{-1}|xu\phi_4|^2 + \tilde{b}\bar{\phi}_4^{-1}(x\bar{v}_4\bar{\phi}_4)^2)dx\end{aligned} \tag{3.16}$$

which gives the following estimate:

$$\frac{d}{dt}\int \Big[\tilde{a}|xv_4\phi_4|^2 + \mathrm{Re}(\tilde{b}\bar{\phi}_4\phi_4^{-1}(x\bar{v}_4\phi_4)^2)\Big]dx \le C(\|xv_4\phi_4\|_{L^2} + \|xv_4\phi_4\|_{L^2}^2). \tag{3.17}$$

With this estimate, we consider the expression

$$\begin{aligned}|||\cdot|||_*^2 = &\sum_{j=0}^{k-1}\|v_j\|_{L^2}^2 + \int \tilde{a}|v_k\phi|^2 + \mathrm{Re}(\tilde{b}\bar{\phi}\phi^{-1}(\bar{v}_k\phi)^2)dx \\ &+ \sum_{j=0}^{3}\|xv_j\|_{L^2}^2 + \int \tilde{a}|xv_4\phi_4|^2 + \mathrm{Re}(\tilde{b}\bar{\phi}_4\phi_4^{-1}(x\bar{v}_4\phi_4)^2)dx.\end{aligned} \tag{3.18}$$

Thus we have that

$$\frac{d}{dt}|||v|||_* \le C(|||v|||_* + |||v|||_*^l)(e^{C(|||v|||_* + |||v|||_*^l)} + 1) \tag{3.19}$$

which shows there exists $T_* > 0$ such that

$$|||v(t)|||_* \le 100|||v(0)|||_* \tag{3.20}$$

and within this range

$$|||\cdot|||_* \sim \sum_{j=0}^{k}\|\partial_x^j u\|_{L^2} + \sum_{j=0}^{4}\|x\partial_x^j u\|_{L^2}. \tag{3.21}$$

Once we obtain this a priori estimate, it is easier to return to the equation (assuming the existence of the solution) to prove that the solution belongs to the space in the statement in Theorem 2. To establish local existence, we again introduce two viscosity terms and proceed exactly as in sections 2.3 and 2.4 with the same argument as above when dealing with gauge transformation for the case $j = k$.

The rest of the proof follows by the method in the previous proof, in the previous section, thus it will be omitted here.

4. Appendix. We would use that for $f, g \in H^r(\mathbb{R})$, $r \ge 1$, then $fg \in H^r(\mathbb{R})$ (i.e., $H^r(\mathbb{R})$ is a Banach algebra under the pointwise product) and

$$\|fg\|_{H^r} \le C_r(\|f\|_{L^\infty}\|g\|_{H^r} + \|g\|_{L^\infty}\|f\|_{H^r}) \tag{4.1}$$

for some $C_r > 0$.

4.1. Appendix 1. Consider the jth derivative of (2.1):

$$\begin{aligned}\partial_x^j\partial_t u &= \partial_x^j\big(ia\partial_x^2 u + ib\partial_x^2\bar{u} + c\partial_x u + d\partial_x\bar{u} + f\big)\\ &= i\partial_x^j(a\partial_x^2 u) + i\partial_x^j(b\partial_x^2\bar{u}) + \partial_x^j(c\partial_x u) + \partial_x^j(d\partial_x\bar{u}) + \partial_x^j f.\end{aligned} \tag{4.2}$$

Write $\partial_x a = a'$, $\frac{\partial}{\partial z_l}a(z_1,z_2,z_3,z_4) = a_l'$, $l = 1,2,3,4$, similarly for b, c, d.

In order to rewrite the jth derivative of (2.1) such that the energy estimate can be evaluated effectively, the terms involving the highest and the second highest derivatives are essential. Thus we need the following:

$$\begin{aligned}\partial_x^j(ia\partial_x^2 u) = \partial_x^j(iav_2) &= i\sum_{l=0}^{j}\binom{j}{l}(\partial_x^l a)(\partial_x^{j-l}v_2)\\ &= ia\partial_x^2 v_j + ija'\partial_x v_j + i(\partial_x^j a)v_2 + i\sum_{l=2}^{j-1}\binom{j}{l}(\partial_x^l a)(\partial_x^{j-l}v_2).\end{aligned} \tag{4.3}$$

Examine $\partial_x^j a$:

$$\begin{aligned}\partial_x^j a = \partial_x^{j-1}(\partial_x a) &= \partial_x^{j-1}(a_1'\partial_x u + a_2'\partial_x\bar{u} + a_3'\partial_x^2 u + a_1'\partial_x^2\bar{u})\\ &= \partial_x^{j-1}(a_1'v_1 + a_2'\bar{v}_1) + a_3'\partial_x v_j + a_4'\partial_x\bar{v}_j\\ &\quad + \sum_{l=1}^{j-1}\binom{j-1}{l}\big((\partial_x^l(a_3'))(\partial_x^{j-1-l}v_2) + (\partial_x^l(a_4'))(\partial_x^{j-1-l}\bar{v}_2)\big),\end{aligned} \tag{4.4}$$

and thus

$$\begin{aligned}\partial_x^j(ia\partial_x^2 u) &= ia\partial_x^2 v_j + ija'\partial_x v_j + ia_3'v_2\partial_x v_j + ia_4'v_2\partial_x\bar{v}_j\\ &\quad + iv_2\partial_x^{j-1}(a_1'v_1 + a_2'\bar{v}_1) + i\sum_{l=2}^{j-1}\binom{j}{l}(\partial_x^l a)v_{j+2-l}\\ &\quad + i\sum_{l=1}^{j-1}\binom{j-1}{l}v_2\big((\partial_x^l(a_3'))v_{j+1-l} + (\partial_x^l(a_4'))\bar{v}_{j+1-l}\big),\end{aligned} \tag{4.5}$$

similarly for $\partial_x^j(ib\partial_x^2\bar{u})$, $\partial_x^j(c\partial_x u)$, $\partial_x^j(d\partial_x\bar{u})$. Write

$$\begin{cases}\partial_x^j(ia\partial_x^2 u) &= ia\partial_x^2 v_j + ija'\partial_x v_j + ia_3'v_2\partial_x v_j + ia_4'v_2\partial_x\bar{v}_j + f_{aj},\\ \partial_x^j(ib\partial_x^2\bar{u}) &= ib\partial_x^2\bar{v}_j + ijb'\partial_x\bar{v}_j + ib_3'\bar{v}_2\partial_x v_j + ib_4'\bar{v}_2\partial_x\bar{v}_j + f_{bj},\\ \partial_x^j(c\partial_x u) &= c\partial_x v_j + c_3'v_1\partial_x v_j + c_4'v_1\partial_x\bar{v}_j + f_{cj},\\ \partial_x^j(d\partial_x\bar{u}) &= d\partial_x\bar{v}_j + d_3'\bar{v}_1\partial_x v_j + d_4'\bar{v}_1\partial_x\bar{v}_j + f_{dj},\end{cases} \tag{4.6}$$

where

$$
\begin{aligned}
f_{aj} &= iv_2\partial_x^{j-1}(a_1'v_1 + a_2'\bar v_1) + i\sum_{l=2}^{j-1}\binom{j}{l}(\partial_x^l a)v_{j+2-l}\\
&\quad + i\sum_{l=1}^{j-1}\binom{j-1}{l}v_2\big((\partial_x^l(a_3'))v_{j+1-l} + (\partial_x^l(a_4'))\bar v_{j+1-l}\big),\\
f_{bj} &= i\bar v_2\partial_x^{j-1}(b_1'v_1 + b_2'\bar v_1) + i\sum_{l=2}^{j-1}\binom{j}{l}(\partial_x^l b)\bar v_{j+2-l}\\
&\quad + i\sum_{l=1}^{j-1}\binom{j-1}{l}\bar v_2\big((\partial_x^l(b_3'))v_{j+1-l} + (\partial_x^l(b_4'))\bar v_{j+1-l}\big),\\
f_{cj} &= v_1\partial_x^{j-1}(c_1'v_1 + c_2'\bar v_1) + \sum_{l=1}^{j-1}\binom{j}{l}(\partial_x^l c)\bar v_{j+1-l}\\
&\quad + \sum_{l=1}^{j-1}\binom{j-1}{l}\big((\partial_x^l c_3')v_{j+1-l} + (\partial_x^l c_4')\bar v_{j+1-l}\big)v_1,\\
f_{dj} &= \bar v_1\partial_x^{j-1}(d_1'v_1 + d_2'\bar v_1) + \sum_{l=1}^{j-1}\binom{j}{l}(\partial_x^l d)\bar v_{j+1-l}\\
&\quad + \sum_{l=1}^{j-1}\binom{j-1}{l}\big((\partial_x^l d_3')v_{j+1-l} + (\partial_x^l d_4')\bar v_{j+1-l}\big)\bar v_1.
\end{aligned}
\tag{4.7}
$$

Combine all of the above and write the jth derivative of (2.1)

$$
\begin{aligned}
\partial_t v_j &= \partial_x^j(ia\partial_x^2 u) + \partial_x^j(ib\partial_x^2\bar u) + \partial_x^j(c\partial_x u) + \partial_x^j(d\partial_x\bar u) + \partial_x^j f\\
&= ia\partial_x^2 v_j + ija'\partial_x v_j + ia_3'v_2\partial_x v_j + ia_4'v_2\partial_x\bar v_j + f_{aj}\\
&\quad + ib\partial_x^2\bar v_j + ijb'\partial_x\bar v_j + ib_3'\bar v_2\partial_x v_j + ib_4'\bar v_2\partial_x\bar v_j + f_{bj}\\
&\quad + c\partial_x v_j + c_3'v_1\partial_x v_j + c_4'v_1\partial_x\bar v_j + f_{cj}\\
&\quad + d\partial_x\bar v_j + d_3'\bar v_1\partial_x v_j + d_4'\bar v_1\partial_x\bar v_j + f_{dj} + \partial_x^j f\\
&= ia\partial_x^2 v_j + ib\partial_x^2\bar v_j + c_j\partial_x v_j + d_j\partial_x\bar v_j + f_j,
\end{aligned}
\tag{4.8}
$$

where

$$
\begin{cases}
c_j = ija' + ia_3'v_2 + ib_3'\bar v_2 + c + c_3'v_1 + d_3'\bar v_1,\\
d_j = ia_4'v_2 + ijb' + ib_4'\bar v_2 + c_4'v_1 + d + d_4'\bar v_1,\\
f_j = f_{aj} + f_{bj} + f_{cj} + f_{dj} + \partial_x^j f.
\end{cases}
\tag{4.9}
$$

4.2. Appendix 2. Let $P(\alpha; z_1, z_2, \dots, z_n)$ represent the polynomial of α-degree in $z_1, z_2, \dots, z_n$.

In order to obtain the estimates (2.6) and (2.22), we need to examine the terms involving a, its jth derivative with respect to x, its derivative with respect to t, and its derivatives with respect to its arguments, e.g., a_l', $l = 1, 2, 3, 4$ etc., similarly for b, c, d and their conjugates. Also, since the expressions $f_{aj}, f_{bj}, f_{cj}, f_{dj}$ in (4.7), F_k, and their conjugates involve the terms mentioned above, they can be considered in

the similar manner. Other terms worth mentioning are α_{12}^k, ϕ and its derivatives with respect to x and t.

Recall that

$$\begin{cases} \tilde{a} = \dfrac{a}{a^2-|b|^2}, \quad \tilde{b} = \dfrac{b}{a^2-|b|^2}, \\ c_j = ija' + ia_3'v_2 + ib_3'\bar{v}_2 + c + c_3'v_1 + d_3'\bar{v}_1, \\ d_j = ijb' + ib_4'v_2 + ib_4'\bar{v}_2 + d + c_4'v_1 + d_4'\bar{v}_1, \\ \alpha_{12}^k = \tilde{a}d_k + \tilde{b}\bar{c}_k, \\ f_j = f_{aj} + f_{bj} + f_{cj} + f_{dj} + \partial_x^j f, \\ F_k = \tilde{a}f_k + \tilde{b}\bar{f}_k. \end{cases} \tag{4.10}$$

We now investigate the terms above in more detail, especially the order with respect to u, $\bar{u}$ and their derivatives. It is enough to look at a and its derivatives; the cases for b, c, d are similar.

$$\begin{aligned} a' &= a_1'\partial_x u + a_2'\partial_x\bar{u} + a_3'\partial_x^2 u + a_4'\partial_x^2\bar{u} \\ &= P(1;\partial_x u, \partial_x\bar{u}, \partial_x^2 u, \partial_x^2\bar{u}), \\ a'' &= P(2;\partial_x u, \partial_x\bar{u}, \partial_x^2 u, \partial_x^2\bar{u}, \partial_x^3 u, \partial_x^3\bar{u}), \\ \partial_x\tilde{a} &= \partial_x\Big(\frac{a}{a^2-|b|^2}\Big) = \frac{a'(a^2-|b|^2) - a(2aa' - \bar{b}b' - b\bar{b}')}{(a^2-|b|^2)^2} \\ &= P(1;u,\bar{u},\dots,\partial_x^2 u, \partial_x^2\bar{u}), \\ \partial_t a &= a_1'\partial_t u + a_2'\partial_t\bar{u} + a_3'\partial_x\partial_t u + a_4'\partial_x\partial_t\bar{u} \\ &= a_1'(ia\partial_x^2 u + ib\partial_x^2\bar{u} + c\partial_x u + d\partial_x\bar{u} + f) \\ &\quad + a_2'(-ia\partial_x^2\bar{u} - i\bar{b}\partial_x^2 u + \bar{c}\partial_x\bar{u} + \bar{d}\partial_x u + \bar{f}) \\ &\quad + a_3'\partial_x(ia\partial_x^2 u + ib\partial_x^2\bar{u} + c\partial_x u + d\partial_x\bar{u} + f) \\ &\quad + a_4'\partial_x(-ia\partial_x^2\bar{u} - i\bar{b}\partial_x^2 u + \bar{c}\partial_x\bar{u} + \bar{d}\partial_x u + \bar{f}) \\ &= P(2;u,\bar{u},\partial_x u,\partial_x\bar{u},\partial_x^2 u,\partial_x^2\bar{u},\partial_x^3 u,\partial_x^3\bar{u}) \\ \partial_t\tilde{a} &= P(2;u,\bar{u},\partial_x u,\partial_x\bar{u},\partial_x^2 u,\partial_x^2\bar{u},\partial_x^3 u,\partial_x^3\bar{u}). \end{aligned} \tag{4.11}$$

Other expressions worth mentioning are c_j, d_j and their derivatives. Again, it is enough to consider c_j:

$$\begin{aligned} \partial_t c_j &= \partial_t(ija' + ia_3'v_2 + ib_3'\bar{v}_2 + c + c_3'v_1 + d_3'\bar{v}_1) \\ &= ij\partial_x\partial_t a + i\partial_t(a_3')\partial_x^2 u + a_3'\partial_x^2\partial_t u + i\partial_t(b_3')\partial_x^2\bar{u} + b_3'\partial_x^2\partial_t\bar{u} \\ &\quad + \partial_t c + \partial_t(c_3')\partial_x u + c_3'\partial_x\partial_t u + \partial_t(d_3')\partial_x\bar{u} + d_3'\partial_x\partial_t\bar{u} \\ &= P(3;u,\bar{u},\dots,\partial_x^4 u,\partial_x^4\bar{u}). \end{aligned} \tag{4.12}$$

We would also consider α_{12}^k since it is an important component of ϕ:

$$\begin{cases} \alpha_{12}^k = \tilde{a}c_k + \tilde{b}d_k = \dfrac{ac_k + b\bar{d}_k}{a^2-|b|^2} = P(u,\bar{u},\partial_x u,\partial_x\bar{u},\partial_x^2 u,\partial_x^2\bar{u}), \\ \partial_x\alpha_{12}^k = d_k\partial_x\tilde{a} + \tilde{a}\partial_x d_k + \bar{c}_k\partial_x\tilde{b} + \tilde{b}\partial_x\bar{c}_k = P(2;u,\bar{u},\dots,\partial_x^3 u,\partial_x^3\bar{u}). \end{cases} \tag{4.13}$$

The boundedness of ϕ and its derivatives play an important role in the energy estimates, thus a closer look at the terms involved is necessary:

$$\begin{aligned}
\phi =& \exp\left\{\frac{1}{2i}\int_0^x (\tilde{a}c_k+\tilde{b}d_k)(x',t)dx'\right\},\\
\partial_x\phi =& \frac{1}{2i}\phi(\tilde{a}c_k+\tilde{b}\bar{d}_k) = P(u,\bar{u},\dots,\partial_x^2 u,\partial_x^2\bar{u}),\\
\partial_x^2\phi =& -\frac{1}{4}\phi(\tilde{a}c_k+\tilde{b}\bar{d}_k)^2+\frac{1}{2i}\phi\partial_x(\tilde{a}c_k+\tilde{b}\bar{d}_k) = P(u,\bar{u},\dots,\partial_x^3 u,\partial_x^3\bar{u}),\\
\partial_t\phi =& \phi\left(\frac{1}{2i}\int_0^x \partial_t\left(\frac{ac_k+b\bar{d}_k}{a^2-|b|^2}\right)(x',t)dx'\right) = P(u,\bar{u},\dots,\partial_x^4 u,\partial_x^4\bar{u});
\end{aligned} \tag{4.14}$$

higher derivatives of ϕ can be derived similarly.

4.2.1. Estimate (2.22). To obtain estimate (2.22), i.e., the boundedness of J_1, we need the following:

$$\begin{aligned}
\|\phi(t)\|_{L^\infty} &\le \exp\left\{\frac{1}{2}\int_{\mathbb{R}}|(\tilde{a}c_k-\tilde{b}d_k)(x,t)|dx\right\}\\
&\le \exp\left\{C\int |ac_k|+|bd_k|dx\right\}\\
&\le \exp\left\{C(\|u\|_{H^k}^2+\|u\|_{H^k}^l)(t)\right\}
\end{aligned} \tag{4.15}$$

for some constant $C>0$, where a,b,c,d are not linear and l depends on the highest order of ac_k+bd_k.

Since $F_k\bar{v}_k=(\tilde{a}f_k+\tilde{b}\bar{f}_k)\bar{v}_k$ is not linear and involves only at most $v_k,\bar{v}_k$, the right-hand side of (2.22) is

$$\begin{aligned}
&2\Big\{\|\partial_x\big(\alpha_{12}^k\phi\bar{\phi}^{-1}\big)\|_{L^\infty}+\|\partial_t\phi\tilde{a}\phi^{-1}\|_{L^\infty}+\|\partial_t\phi\tilde{b}\bar{\phi}^{-1}\|_{L^\infty}\\
&\quad+\|\alpha_{12}^k\phi\bar{\phi}^{-2}\partial_x\bar{\phi}\|_{L^\infty}+\|(\partial_x^2\phi)\phi^{-1}\|_{L^\infty}+\|\partial_t\left(b\bar{\phi}\phi^{-1}\right)\|_{L^\infty}\\
&\quad+\frac{1}{2}\|\partial_t\tilde{a}\|_{L^\infty}\Big\}\|v_k\phi\|_{L^2}^2+2\int|\phi F_k\bar{v}_k\bar{\phi}|dx\le J_1\|v_k\|_{L^2}^2
\end{aligned} \tag{4.16}$$

with

$$J_1\le e^{C(\|u\|_{H^k}^2+\|u\|_{H^k}^l)}\Big(\|u\|_{H^k}^2+\|u\|_{H^k}^{l-1}\Big). \tag{4.17}$$

4.3. Appendix 3.

4.3.1. Proof of Proposition 2.1. The proof of Proposition 2.1 is based on the following equations:

$$\phi\partial_x^2 v_k=\partial_x^2(v_k\phi)-2\partial_x\phi\phi^{-1}\partial_x(v_k\phi)+[2(\partial_x\phi)^2\phi^{-2}-\partial_x^2\phi\phi^{-1}](v_k\phi), \tag{4.18}$$

$$\begin{aligned}
\phi\partial_x^3 v_k=&\partial_x^3(v_k\phi)-3\partial_x\phi\phi^{-1}\partial_x^2(v_k\phi)+[6(\partial_x\phi)^2\phi^{-1}-3\partial_x^2\phi]\phi^{-1}\partial_x(v_k\phi)\\
&+(-(6(\partial_x\phi)^2\phi^{-1}-3\partial_x^2\phi)\phi^{-1}\partial_x\phi-\partial_x^3\phi)\phi^{-1}(v_k\phi),
\end{aligned} \tag{4.19}$$

$$\begin{aligned}
\phi\partial_x^4 v_k=&\partial_x^4(v_k\phi)-4\partial_x\phi\phi^{-1}\partial_x^3(v_k\phi)+[12(\partial_x\phi)^2\phi^{-1}-6\partial_x^2\phi]\phi^{-1}\partial_x^2(v_k\phi)\\
&+[-24(\partial_x\phi)^3\phi^{-1}+24\phi^{-1}\partial_x^2\phi\partial_x\phi-4\partial_x^3\phi]\phi^{-1}\partial_x(v_k\phi)\\
&+\{24(\partial_x\phi)^4\phi^{-2}-12[2\phi^{-2}+\phi^{-1}]\partial_x^2\phi(\partial_x\phi)^2+8\phi^{-1}\partial_x\phi\partial_x^3\phi\\
&+6\phi^{-1}(\partial_x^2\phi)^2-\partial_x^4\phi\}\phi^{-1}(v_k\phi)\\
=&\partial_x^4(v_k\phi)+\gamma_3\partial_x^3(v_k\phi)+\gamma_2\partial_x^2(v_k\phi)+\gamma_1\partial_x(v_k\phi)+\gamma_0 v_k\phi.
\end{aligned} \tag{4.20}$$

We would want to show that for some choice of $\delta > 0$ we are able to bound the following four terms within $H^k(\mathbb{R})$:

(4.21)
$$2\mathrm{Re}\int(\bar{v}_k\bar{\phi})\tilde{a}(-\varepsilon\phi\partial_x^4 v_k+\varepsilon\delta\phi\partial_x^2 v_k)dx+2\mathrm{Re}\int(\bar{v}_k\bar{\phi})\tilde{b}(-\varepsilon\phi\partial_x^4\bar{v}_k+\varepsilon\delta\phi\partial_x^2\bar{v}_k)dx.$$

Using (4.18)–(4.20) and after integration by parts, we have

$$\begin{aligned}
&2\mathrm{Re}\int(\bar{v}_k\bar{\phi})\tilde{a}(-\varepsilon\phi\partial_x^4 v_k+\varepsilon\delta\phi\partial_x^2 v_k)dx\\
&\qquad+2\mathrm{Re}\int(\bar{v}_k\bar{\phi})\tilde{b}(-\varepsilon\phi\partial_x^4\bar{v}_k+\varepsilon\delta\phi\partial_x^2\bar{v}_k)dx\\
&\le-2\varepsilon\int(\tilde{a}-|\tilde{b}|)|\partial_x^2(v_k\phi)|^2dx-2\varepsilon\delta\int(\tilde{a}-|\tilde{b}|)|\partial_x(v_k\phi)|^2dx\\
&\quad+\varepsilon C_1\|\partial_x^2(v_k\phi)\|_{L^2}\|\partial_x(v_k\phi)\|_{L^2}+\varepsilon C_2\|\partial_x(v_k\phi)\|_{L^2}\|v_k\phi\|_{L^2}\\
&\quad+\varepsilon C_3\|\partial_x(v_k\phi)\|_{L^2}^2+\varepsilon C_4\|v_k\phi\|_{L^2}^2.
\end{aligned}\tag{4.22}$$

Thus for $p,q>0$, we have

$$\begin{aligned}
&2\mathrm{Re}\int(\bar{v}_k\bar{\phi})\tilde{a}(-\varepsilon\phi\partial_x^4 v_k+\varepsilon\delta\phi\partial_x^2 v_k)dx\\
&\qquad+2\mathrm{Re}\int(\bar{v}_k\bar{\phi})\tilde{b}(-\varepsilon\phi\partial_x^4\bar{v}_k+\varepsilon\delta\phi\partial_x^2\bar{v}_k)dx\\
&\le-2\varepsilon m_2\|\partial_x^2(v_k\phi)\|_{L^2}^2-2\varepsilon\delta m_2\|\partial_x(v_k\phi)\|_{L^2}^2\\
&\quad+\varepsilon C_1\left(\frac{1}{p}\|\partial_x^2(v_k\phi)\|_{L^2}^2+p\|\partial_x(v_k\phi)\|_{L^2}^2\right)\\
&\quad+\varepsilon C_2\left(\frac{1}{q}\|\partial_x(v_k\phi)\|_{L^2}^2+q\|v_k\phi\|_{L^2}^2\right)+\varepsilon C_3\|\partial_x(v_k\phi)\|_{L^2}^2+\varepsilon C_4\|v_k\phi\|_{L^2}^2\\
&=\varepsilon\left(-2m_2+C_1p^{-1}\right)\|\partial_x^2(v_k\phi)\|_{L^2}^2\\
&\quad+\varepsilon\left(-2\delta c_3+C_1p+C_2q^{-1}+C_3\right)\|\partial_x(v_k\phi)\|_{L^2}^2+\varepsilon(C_2q+\delta C_4)\|v_k\phi\|_{L^2}^2,
\end{aligned}\tag{4.23}$$

where

$$\begin{aligned}
C_1&=2\|\tilde{a}\gamma_2\|_{L^\infty},\\
C_2&=\|\partial_x^2(\tilde{a}\gamma_3)\|_{L^\infty}+2\varepsilon\|\partial_x(\tilde{a}\gamma_2)\|_{L^\infty}+2\varepsilon\|\tilde{a}\gamma_1\|_{L^\infty}+4\varepsilon\delta\|\tilde{a}\partial_x\phi\phi^{-1}\|_{L^\infty},\\
C_3&=4\|\partial_x^2\tilde{a}\|_{L^\infty}+2\|\partial_x(\tilde{a}\gamma_3)\|_{L^\infty}+2\|\tilde{a}\gamma_2\|_{L^\infty}\\
&\quad+2\|\tilde{b}\bar{\phi}^{-1}\phi\bar{\gamma}_2\|_{L^\infty}+4\|\partial_x^2(\tilde{b}\bar{\phi}^{-1}\phi)\|_{L^\infty},\\
C_4&=\|\partial_x^4\tilde{a}\|_{L^\infty}+2\|\tilde{a}\gamma_0\|_{L^\infty}+2\delta\|\tilde{a}[2(\partial_x\phi)^2\phi^{-2}-\partial_x^2\phi\phi^{-1}]\|_{L^\infty}\\
&\quad+2\|\partial_x^4(\tilde{b}\bar{\phi}^{-1}\phi)\|_{L^\infty}+2\|\partial_x^3(\tilde{b}\bar{\phi}^{-1}\phi\bar{\gamma}_3)\|_{L^\infty}+\|\partial_x^2(\tilde{b}\bar{\phi}^{-1}\phi\bar{\gamma}_2)\|_{L^\infty}\\
&\quad+2\|\partial_x(\tilde{b}\bar{\phi}^{-1}\phi\bar{\gamma}_1)\|_{L^\infty}+2\|\tilde{b}\bar{\phi}^{-1}\phi\bar{\gamma}_2\|_{L^\infty}+2\delta\Big[\|\partial_x^2(\tilde{b}\bar{\phi}^{-1}\phi)\|_{L^\infty}\\
&\quad+\|\partial_x(\tilde{b}\bar{\phi}^{-2}\phi\partial_x\bar{\phi})\|_{L^\infty}+\|\tilde{b}(2\partial_x\bar{\phi})^2\bar{\phi}^{-2}-(\partial_x^2\bar{\phi})\bar{\phi}^{-1}\|_{L^\infty}\Big].
\end{aligned}\tag{4.24}$$

To obtain the desired result in Proposition 2.1, we choose p such that $2m_2 \geq C_1 p^{-1}$, or, $p \geq \frac{C_1}{2m_2}$, and then choose δ such that $-2\delta m_2 + C_1 p + C_2 q^{-1} + C_3 \leq 0$, or $\delta \geq \frac{C_1 p + C_2 q^{-1} + C_3}{2m_2}$. Thus

$$2\mathrm{Re}\int (\bar{v}_k\bar{\phi})\tilde{a}(-\varepsilon\phi\partial_x^4 v_k + \varepsilon\delta\phi\partial_x^2 v_k)dx + 2\mathrm{Re}\int (\bar{v}_k\bar{\phi})\tilde{b}(-\varepsilon\phi\partial_x^4 \bar{v}_k + \varepsilon\delta\phi\partial_x^2 \bar{v}_k)dx \leq C\|v_k\phi\|_{L^2}^2 \tag{4.25}$$

for $\varepsilon \in (0,1]$, where $C = C(\delta, \|u\|_{H^k}) > 0$ is independent of ε.

Acknowledgment. The authors would like to thank the referees for several suggestions which improved the presentation of this work.

REFERENCES

[1] A. de Bouard, N. Hayashi, P.O. Naumkin, and J.-C. Saut, *Scattering problem and asymptotics for a relativistic nonlinear Schrödinger equation*, Nonlinearity, 12 (1999), pp. 1415–1425.

[2] A. de Bouard, N. Hayashi, and J.-C. Saut, *Global existence of small solutions to a relativistic nonlinear Schrödinger equation*, Comm. Math. Phys., 189 (1997), pp. 73–105.

[3] H. Chihara, *Local existence for semilinear Schrödinger equations*, Math. Japon., 42 (1995), pp. 35–51.

[4] H. Chihara, *The initial value problem for semilinear Schrödinger equations*, Publ. Res. Inst. Math. Sci., 32 (1996), pp. 445–471.

[5] M. Colin, Ph.D. thesis, Université de Paris-Sud, Orsay Cedex, France.

[6] M. Colin, *On the local well-posedness of quasilinear Schrödinger equations in arbitrary space dimension*, Comm. Partial Differential Equations, 27 (2002), pp. 325–354.

[7] P. Constantin and J.-C. Saut, *Local smoothing properties of dispersive equations*, J. Amer. Math. Soc., 1 (1989), pp. 413–446.

[8] S. Doi, *On the Cauchy problem for Schrödinger type equations and the regularity of the solutions*, J. Math. Kyoto Univ., 34 (1994), pp. 319–328.

[9] S. Doi, *Remarks on the Cauchy problem for Schrödinger type equations*, Comm. Partial Differential Equations, 21 (1996), pp. 163–178.

[10] N. Hayashi and E.I. Kaikina, *Local existence of solutions to the Cauchy problem for nonlinear Schrödinger equations*, SUT J. Math., 34 (1998), pp. 111–137.

[11] N. Hayashi and T. Ozawa, *Remarks on Schrödinger equations in one space dimension*, Differential Integral Equations, 2 (1994), pp. 453–461.

[12] T. Kato, *Nonlinear Schrödinger equations*, in Schrödinger Operators, Lecture Notes in Phys. 345, H. Holden and A. Jensen, eds., Springer-Verlag, Berlin, New York, 1989, pp. 218–263.

[13] C.E. Kenig, G. Ponce, and L. Vega, *Oscillatory integrals and regularity of dispersive equations*, Indiana Univ. Math. J., 40 (1991), pp. 33–69.

[14] C.E. Kenig, G. Ponce, and L. Vega, *Small solutions to nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 255–288.

[15] C.E. Kenig, G. Ponce, C. Rolvung, and L. Vega, *Local existence theory for the generalized nonlinear Schrödinger equations*, to appear.

[16] S. Mizohata, *On the Cauchy Problem*, Notes Rep. Math. Sci. Engrg. 3, Science Press, Beijing, Academic Press, New York, 1985.

[17] T. Ozawa, *Remarks on quadratic nonlinear Schrödinger equations*, Funkcial. Ekvac., 38 (1995), pp. 217–232.

[18] M. Poppenberg, *On the local wellposedness for quasiliniear Schrödinger equations in arbitrary space dimension*, J. Differential Equations, 172 (2001), pp. 83–115.

[19] M. Poppenberg, *Smooth solutions for a class of fully nonlinear Schrödinger type equations*, Nonlinear Anal., 45 (2001), pp. 723–741.

[20] P. Sjölin, *Regularity of solutions to the Schrödinger equations*, Duke Math. J., 55 (1987), pp. 699–715.

[21] L. Vega, *The Schrödinger equation: Pointwise convergence to the initial data*, Proc. Amer. Math. Soc., 102 (1998), pp. 874–878.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 460–477

A DUALITY APPROACH FOR THE BOUNDARY VARIATION OF NEUMANN PROBLEMS*

DORIN BUCUR† AND NICOLAS VARCHON‡

Abstract. In two dimensions, we study the stability of the solution of an elliptic equation with Neumann boundary conditions for nonsmooth perturbations of the geometric domain. Using harmonic conjugates, we relate this problem to the shape stability of the solution of an elliptic equation with Dirichlet boundary conditions. As a particular case, we prove the stability of the solution under a topological constraint (uniform number of holes), which is analogous to Šverák's result for Dirichlet boundary conditions.

Key words. boundary variation, Neumann problem, shape optimization, stability

AMS subject classifications. 35J20, 35B20

PII. S0036141002389579

1. Introduction.

An interesting question arising in shape optimization concerns the stability of the solution of a partial differential equation (PDE) for nonsmooth variations of the geometric domain. Various papers in the literature deal with PDEs with Dirichlet boundary conditions, while very few results can be found for PDEs with Neumann boundary conditions. Several reasons may explain this situation, but maybe the most important is that for a nonsmooth open set Ω a function of the Sobolev space $H^1(\Omega)$ might not have an extension outside Ω.

The purpose of this paper is to give a quite general method, based on duality, for the study of the shape stability of the weak solution of a linear elliptic problem with homogeneous Neumann boundary conditions. Given two bounded open sets $\Omega \subseteq D \subseteq \mathbb{R}^2$, $a \in L^\infty(D), a \geq 0$, and $h \in L^2(D)$, we consider the following problem:

$$\left\{ \begin{array}{rl} -\triangle u_{\Omega,h} + a(x) u_{\Omega,h} = h & \text{in} \quad \Omega, \\ \frac{\partial u_{\Omega,h}}{\partial n} = 0 & \text{on} \quad \partial\Omega. \end{array} \right. \tag{1}$$

In order to handle easily the compatibility conditions on regions where a vanishes, we suppose that $h(x) = a(x)f(x) + g(x)$, where $f \in L^2(D)$ and $g \in L^2(D)$, supp $g \subseteq \Omega$, and $\int_C g dx = 0$ for every connected component C of Ω.

We study the stability of the solution $u_{\Omega,h}$ for perturbations of the geometric domain Ω inside D, i.e., the "continuity" of the mapping $\Omega \mapsto u_{\Omega,h}$. We point out that we consider only *weak* solutions of (1) (see the precise definition in section 2); those solutions are classical only if Ω, a, and h are regular enough.

The family of domains is endowed with the Hausdorff complementary topology (see [7, 24, 25]), which has good compactness properties and allows nonsmooth perturbations of the boundaries. We are particularly interested in dealing with nonsmooth domains, like domains with cracks or with boundaries of strictly positive measure. For this reason, the functional spaces where the weak solutions are defined play a

*Received by the editors January 31, 2002; accepted for publication (in revised form) May 1, 2002; published electronically December 3, 2002.
http://www.siam.org/journals/sima/34-2/38957.html
†Département de Mathématiques, Université de Metz, Ile du Saulcy, 57045 Metz, France (bucur@poncelet.univ-metz.fr).
‡Department of Mathematics, Technical University of Denmark, Building 303, DK-2800 Lyngby, Denmark (N.Varchon@mat.dtu.dk).

crucial role; one has to pay attention to the fact that extension operators may fail to exist as soon as Ω is not smooth (for example, if Ω has a crack, there is no extension operator from $H^1(\Omega)$ to $H^1(\mathbb{R}^2)$).

In order to compare two solutions on two different domains, the following convention is applied: extending by zero on $D\setminus\Omega$, we see $u_{\Omega,h}$ and $\nabla u_{\Omega,h}$ as functions defined on D. The exact sense of those extensions is given in section 2 once the functional spaces where the solutions belong are introduced.

Several results can be found in the literature concerning (1) for $a \equiv 1$, which has a variational solution in the Sobolev space $H^1(\Omega)$. We refer to [10] for a pioneering continuity result obtained under geometric constraints on the variable domains (uniform Lipschitz boundary), which particularly imply the existence of uniformly bounded extension operators from $H^1(\Omega)$ to $H^1(\mathbb{R}^2)$; the existence of extension operators across the boundary is the key result for the shape continuity. In [23] the shape continuity is established for the same equation in a class of domains satisfying weaker geometric constraints which still insure the existence of a dense set of functions having extensions.

A different point of view, still for $a \equiv 1$, based on the Mosco convergence of H^1-spaces, was followed by Chambolle and Doveri in [9]. (In the last section we recall the definition of the Mosco convergence and the main lines of this issue.) Here, the extension property is replaced by an approximability one: the family of functions of $H^1(\Omega)$ which can be written as strong limits of elements of $H^1(\Omega_n)$ is dense in $H^1(\Omega)$. They proved (in two dimensions) that if Ω_n converges in the Hausdorff complementary topology to Ω and the lengths of the boundaries $\mathcal{H}^1(\partial\Omega_n)$ and the number of the connected components of $\partial\Omega_n$ are uniformly bounded, then $u_{\Omega_n,h}$ converges to $u_{\Omega,h}$. In [6] a more general result is proved for the same equation (i.e., $a \equiv 1$): if Ω_n converges in the Hausdorff complementary topology to Ω such that the number of the connected components of $\mathbb{R}^2 \setminus \Omega_n$ is uniformly bounded, then shape continuity holds if and only if the Lebesgue measure is stable, i.e., $|\Omega_n| \to |\Omega|$.

The purpose of this paper is to investigate the case $a \not\equiv 1$ and, in particular, the case when a vanishes on some regions of the plane. For example if $a \equiv 0$, we observe that the stability of the Lebesgue measure is not anymore a necessary condition for the shape stability of the solutions. We give a set of conditions which is equivalent to the shape stability of (1). The major condition, which in concrete examples is the one difficult to check, is reduced by a duality argument to the study of the shape stability of an elliptic equation with Dirichlet boundary conditions. In particular, we prove the following.

THEOREM 1.1. *Let $\{\Omega_n\}_{n\in\mathbb{N}}$ such that $\Omega_n \subseteq D$ and the number of the connected components of Ω_n^c is uniformly bounded. If Ω_n^c converges into the Hausdorff metric to Ω^c, then, for every admissible right-hand side h in* (1)*, we have that $u_{\Omega_n,h}$ converges to $u_{\Omega,h}$ if and only if $|\Omega_n \cap \{a > 0\}| \to |\Omega \cap \{a > 0\}|$.*

The sense of the convergence is defined in section 2. In the extremal case $a \equiv 0$, this result is analogous to the compactness-continuity result of Šverák [25] for Dirichlet boundary conditions.

In section 2 we introduce the main notation. Section 3 is devoted to the case $a \equiv 0$ and to the duality argument. The general case $a \geq 0$ is discussed in section 4. We finish the paper with an example and some remarks.

2. Notation and preliminaries. In this section we set the main notation and recall some facts about the (weak) variational solutions of (1). For this purpose, we introduce the functional spaces in which the solutions are searched.

Let D be a bounded open set in $\mathbb{R}^2$ (called design region). Let $a \in L^\infty(D), a \geq 0$, be a fixed function. For every open set $\Omega \subseteq D$ we introduce the following functional space:

$$\mathcal{L}^{1,2}_a(\Omega) = \left\{ u \in L^2_{loc}(\Omega) : \nabla u \in L^2(\Omega, \mathbb{R}^2), \int_\Omega u^2 a dx < +\infty \right\}, \tag{2}$$

where the gradient of u is taken in the sense of distributions. Introducing the equivalence relation

$$u \mathcal{R}_a v \text{ if } \int_\Omega |\nabla(u-v)|^2 dx + \int_\Omega (u-v)^2 a dx = 0,$$

the quotient space $\mathcal{L}^{1,2}_a(\Omega)_{/\mathcal{R}_a} := L^{1,2}_a(\Omega)$ is a Hilbert space for the scalar product

$$(u,v)_{L^{1,2}_a(\Omega)} = \int_\Omega \nabla u \nabla v dx + \int_\Omega uva dx.$$

Let C be a connected component of Ω and let $u, v \in \mathcal{L}^{1,2}_a(\Omega)$ such that $u \mathcal{R}_a v$. Note that if $|C \cap \{a > 0\}| = 0$, then $u - v$ is constant a.e. on C. If $|C \cap \{a > 0\}| > 0$, this constant is zero, i.e., $u = v$ a.e. on C.

If $a \equiv 1$, then $L^{1,2}_a(\Omega)$ is nothing else but the usual Sobolev space $H^1(\Omega)$ (see [2]). If $a \equiv 0$, then $\mathcal{L}^{1,2}_a(\Omega)$ is the usual Dirichlet space (see [20]). In our paper, if $a \equiv 0$, the spaces $\mathcal{L}^{1,2}_a(\Omega)$, $L^{1,2}_a(\Omega)$ will simply be denoted $\mathcal{L}^{1,2}(\Omega)$, $L^{1,2}(\Omega)$, respectively. Note that if $a_1 \leq a_2$, then the natural injection $L^{1,2}_{a_2}(\Omega) \hookrightarrow L^{1,2}_{a_1}(\Omega)$ is a contraction.

Following [19, Corollary 2.2], if Ω is smooth enough (e.g., with Lipschitz continuous boundary and with a finite number of connected components), then $\mathcal{L}^{1,2}(\Omega) = H^1(\Omega)$. If Ω is not smooth, then $H^1(\Omega)$ might be strictly contained in $\mathcal{L}^{1,2}(\Omega)$. Observe also that if Ω is not smooth enough, several "well-known" properties of H^1-spaces fail to be true, as, for example, the Poincaré–Wirtinger inequality. Moreover, there does not exist an extension operator from $H^1(\Omega)$ to $H^1(D)$, even though the density of $C^\infty(\Omega) \cap H^1(\Omega)$ in $H^1(\Omega)$ remains true (see [18]). In fact, $C^\infty(\overline{\Omega})$ is no longer dense in $H^1(\Omega)$.

Let $h \in L^2(D)$ be such that $h(x) = a(x)f(x) + g(x)$, where $f \in L^2(D)$ and $g \in L^2(D)$, supp $g \subseteq \Omega$, and $\int_C g dx = 0$ for every connected component C of Ω. Then (1) has a weak variational solution $u_{\Omega,h} \in L^{1,2}_a(\Omega)$ obtained by the minimization of the energy functional

$$L^{1,2}_a(\Omega) \ni u \mapsto F(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 dx + \frac{1}{2}\int_\Omega u^2 a dx - \int_\Omega hu dx.$$

This is an immediate consequence of the Lax–Milgram theorem (see [2, Corollary V.8]). The only point to be verified is the strong continuity of the mapping

$$L^{1,2}_a(\Omega) \ni u \mapsto \int_\Omega afu + gu dx.$$

Indeed,

$$\left| \int_\Omega afu + gu dx \right| \leq \int_\Omega |afu| dx + \left| \int_U gu dx \right|$$

$$\leq \Big(\int_\Omega au^2 dx \Big)^{\frac{1}{2}} \Big(\int_\Omega af^2 dx \Big)^{\frac{1}{2}} + C \Big(\int_U |\nabla u|^2 dx \Big)^{\frac{1}{2}} \Big(\int_U g^2 dx \Big)^{\frac{1}{2}} \leq C' |u|_{L^{1,2}_a(\Omega)},$$

where C is the constant given by the Poincaré–Wirtinger inequality applied in $H^1(U)_{/\mathbb{R}}$; the smooth set U is chosen such that supp $g \subseteq \mathrm{U} \subseteq \Omega$.

If Ω, a, and h are smooth enough, every representative in $\mathcal{L}^{1,2}_a(\Omega)$ of the weak variational solution is also classical. In view of the equivalence relation $\mathcal{R}_a$, on each connected component C of Ω, two classical solutions of (1) are identical if $|\{a > 0\} \cap C| > 0$ and differ by a constant if $|\{a > 0\} \cap C| = 0$. It is not our purpose to find the minimal assumptions such that the weak solution is classical (see e.g., [2]); if Ω is of class C^3 and a, h are of class $C^1(\overline{\Omega})$, then every representative of the weak solution is classical. If Ω is not smooth, the sense of the Neumann condition on $\partial\Omega$ is only *weak*; it is implicitly contained in the variational formulation of the problem.

One of the main ideas of this paper is to introduce a new equation which is easier to study from the point of view of the shape stability, but which carries most of the information concerning the shape stability of (1). Let $B = B(0, r)$ be such that $B(0, r + \delta) \subseteq \Omega \subseteq D$ for some $\delta > 0$ and $\gamma \in H^{\frac{1}{2}}(\partial B)$ such that $\int_{\partial B} \gamma d\sigma = 0$. Note that under this last assumption, γ is also an element of the dual of $H^{\frac{1}{2}}(\partial B)_{/\mathbb{R}}$. We consider the following equation:

$$\text{(3)} \qquad \begin{cases} -\triangle v_{\Omega,\gamma} = 0 & \text{in} \quad \Omega \setminus \overline{B}, \\ \frac{\partial v_{\Omega,\gamma}}{\partial n} = 0 & \text{on} \quad \partial\Omega, \\ \frac{\partial v_{\Omega,\gamma}}{\partial n} = \gamma & \text{on} \quad \partial B. \end{cases}$$

Equation (3) has a unique variational solution in $L^{1,2}(\Omega \setminus \overline{B})$ obtained by the minimization of the energy functional

$$\text{(4)} \qquad L^{1,2}(\Omega \setminus \overline{B}) \ni v \mapsto F(v) = \frac{1}{2} \int_{\Omega \setminus \overline{B}} |\nabla v|^2 dx - \int_{\partial B} \gamma v d\sigma.$$

This is a consequence of the Lax–Milgram theorem and, again, the only point to be verified is the continuity of the mapping

$$v \mapsto \int_{\partial B} \gamma v d\sigma.$$

This is a direct consequence of the trace theorem and the Poincaré–Wirtinger inequality applied in $H^1(B(0, r + \delta) \setminus \overline{B})$.

The main interest in relating the shape stability of the solution of (1) to the shape stability of the solution of (3) relies on the fact that all solutions of (3) (even in open sets with nonsmooth boundaries) have harmonic conjugates which satisfy a Dirichlet boundary condition, which is easier to handle on varying domains. Several results for the boundary variation of Dirichlet problems, such as those of [4, 7, 15, 25], can be applied. Observe that a new difficulty (of different type) appears, since the traces of the conjugate functions on the boundary are constant on connected components, but the constants may vary. Nevertheless, in concrete examples, this seems easier to handle, as opposed to directly investigating the stability of the original problem.

The sense in which we investigate the continuity of the mappings

$$\Omega \mapsto u_{\Omega,h} \qquad \text{and} \qquad \Omega \mapsto v_{\Omega,\gamma}$$

is the following. For simplicity, we denote by $L^2_a(D)$ the usual space of square integrable functions with respect to the measure of density $a(x)$ with respect to the

Lebesgue measure, endowed with the scalar product $(u,v)=\int_D uv\ adx$. Since the space to which $u_{\Omega,h}$ belongs varies with Ω, the following convention is used. We embed the space $L^{1,2}_a(\Omega)$ into the following space, which is not dependent on Ω:

$$L^{1,2}_a(\Omega)\hookrightarrow L^2_a(D)\times L^2(D,\mathbb{R}^2) \tag{5}$$

by

$$u\mapsto(\tilde{u},\widetilde{\nabla u}), \tag{6}$$

where $\tilde{u}(x)=u(x)$ if $x\in\Omega$ and $\tilde{u}(x)=0$ if $x\in D\setminus\Omega$. In the same way $\widetilde{\nabla u}(x)=\nabla u(x)$ if $x\in\Omega$ and $\widetilde{\nabla u}(x)=0$ if $x\in D\setminus\Omega$. Note that $\widetilde{\nabla u}$ is not the distributional gradient of $\tilde{u}$.

Since $L^{1,2}_a(\Omega)$ is a quotient space, one has to check that $\tilde{u}$ and $\widetilde{\nabla u}$ do not depend on the choice of the representative of u. This is true since all representatives of u have the same gradient and coincide on $\Omega\cap\{a>0\}$. If $a\equiv 0$, of course the space $L^{1,2}(\Omega)$ is embedded in $L^2(D,\mathbb{R}^2)$, since $L^2_a(\Omega)\equiv\{0\}$.

We denote

$$\mathcal{O}(D)=\{\Omega\subseteq D:\Omega\text{ open}\}\qquad\text{and}\qquad\mathcal{O}_l(D)=\{\Omega\subseteq D:\Omega\text{ open }\sharp\Omega^c\leq l\}.$$

Here $l\in\mathbb{N}$ is fixed, and $\sharp\Omega^c$ denotes the number of the connected components of the complement of Ω.

The Hausdorff distance in the family of open subsets of D (called the Hausdorff complementary distance) is given by the following metric:

$$d_{H^c}(\Omega_1,\Omega_2)=d_H(\overline{D}\setminus\Omega_1,\overline{D}\setminus\Omega_2),$$

where

$$d_H(K_1,K_2)=\max\left\{\sup_{x\in K_1}\inf_{y\in K_2}|x-y|,\sup_{y\in K_2}\inf_{x\in K_1}|x-y|\right\}$$

is the usual Hausdorff distance between two closed sets. It is well known that $\mathcal{O}(D)$ is compact in the Hausdorff complementary topology. Moreover, if $\Omega_n\xrightarrow{H^c}\Omega$, then

$$\forall K\subset\subset\Omega,\ \ \exists n=n_K\in\mathbb{N}\quad\text{such that }\ \forall n\geq n_K\text{ we have }K\subseteq\Omega_n. \tag{7}$$

This property has a geometric character and does not require any regularity on Ω (see [24, Lemma 3, p. 32]). A direct consequence is the following:

$$\forall\phi\in C^\infty_0(\Omega)\quad\exists n=n_\phi\in\mathbb{N}\quad\text{such that }\ \forall n\geq n_\phi\text{ we have }\phi\in C^\infty_0(\Omega_n). \tag{8}$$

The characteristic function of a set E is denoted 1_E and its Lebesgue measure is denoted $|E|$. The capacity of a set E is denoted $cap(E)$; we refer to [20] for details concerning capacity and quasi-continuous functions in Sobolev spaces.

3. The shape stability in the case $a\equiv 0$. In this section we discuss the particular case $a\equiv 0$. We give, in a first step, a proposition relating the shape stability of (1) to the shape stability of (3). In a second step we present the duality method. Using the harmonic conjugates associated to the solutions of (3), we prove the shape stability of (3) for the H^c-topology in the family of domains for which the complements have a uniformly bounded number of connected components.

3.1. Relation between shape stability of (1) with $a \equiv 0$ and (3).

PROPOSITION 3.1. *Let $\Omega_n, \Omega \in \mathcal{O}(D)$ such that $\Omega_n \xrightarrow{H^c} \Omega$. The following assertions are equivalent:*

1. *For $a \equiv 0$ and for every admissible $h := g$, we have $\widetilde{\nabla u_{\Omega_n,g}} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla u_{\Omega,g}}$.*
2. *For every ball B such that $\overline{B} \subseteq \Omega$ and for every $\gamma \in H^{\frac{1}{2}}(\partial B)$ with $\int_{\partial B} \gamma d\sigma = 0$ we have $\widetilde{\nabla v_{\Omega_n,\gamma}} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla v_{\Omega,\gamma}}$,*
3. *For every $u \in L^{1,2}(\Omega)$ there exists $u_n \in L^{1,2}(\Omega_n)$ such that $\widetilde{\nabla u_n} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla u}$.*

We list condition 3 in Proposition 3.1 because it is useful in the proof of the equivalence between conditions 1 and 2.

Proof. $1 \Rightarrow 3$. Let us denote

$$Y = \left\{ \psi \in L^{1,2}(\Omega) : \exists \psi_n \in L^{1,2}(\Omega_n) \quad \text{such that } \widetilde{\nabla \psi_n} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla \psi} \right\}.$$

It is sufficient to prove that Y is dense in $L^{1,2}(\Omega)$; then 3 follows straightforwardly by a usual diagonal procedure. Let $\Psi \in L^{1,2}(\Omega)$ such that

$$\Psi \perp_{L^{1,2}(\Omega)} Y,$$

i.e., $\int_\Omega \nabla\Psi\nabla v dx = 0$ for all $v \in Y$. Let us fix a representative of $\psi \in \mathcal{L}^{1,2}(\Omega)$. Following property (8) of the Hausdorff convergence, the equivalence class generated by $C_0^\infty(\Omega)$ in $L^{1,2}(\Omega)$ is contained in Y. Hence, for all $v \in C_0^\infty(\Omega)$ we have $\int_\Omega \nabla\Psi\nabla v dx = 0$, and therefore $-\Delta\Psi = 0$ in $\mathcal{D}'(\Omega)$. Now let B be a ball such that $\overline{B} \subset \Omega$. For every $g \in L^2(D)$, with supp $g \subset \bar{B}$ and $\int_B g dx = 0$ we have, following condition 1, $\int_B g\Psi dx = 0$, so Ψ is constant in B; hence $\nabla\Psi = 0$ in the connected component of Ω which contains B. Applying this argument to every connected component of Ω, we deduce that $\nabla\Psi = 0$ in Ω, i.e., $\Psi \equiv 0$ in $L^{1,2}(\Omega)$.

$3 \Rightarrow 1$. Let $g \in L^2(\Omega)$ and $K =$ supp g. Let U be a smooth open set such that $K \subset U \subset\subset \Omega$. Taking $u_{\Omega_n,g}$ as a test function in (1) and applying the Poincaré inequality in $H^1(U)$, we obtain that the sequence

$$\|\widetilde{\nabla u_{\Omega_n,g}}\|_{L^2(D,\mathbb{R}^2)}$$

is bounded. Up to a subsequence denoted by the same index we have

$$\widetilde{\nabla u_{\Omega_n,g}} \overset{L^2(D)}{\rightharpoonup} (u_1, u_2).$$

From property (7) of the H^c-convergence, we get that

$$\forall \overline{q} \in C_0^\infty(\Omega, \mathbb{R}^2), \quad \text{div } \overline{q} = 0, \qquad \langle (u_1, u_2)|_\Omega, \overline{q} \rangle_{H^{-1}(\Omega,\mathbb{R}^2)\times H_0^1(\Omega,\mathbb{R}^2)} = 0.$$

Applying successively De Rham's theorem [19, Theorem 2.3] on an increasing sequence of smooth sets covering Ω, there exists $u \in L^2_{loc}(\Omega)$ such that $(u_1, u_2)|_\Omega = \nabla u$ in the distributional sense in Ω. Moreover, from the compact injection $H^1(U) \hookrightarrow L^2(U)$ we have $u_{\Omega_n,g} \overset{L^2(U)}{\longrightarrow} u$. Following condition 3, for every $v \in L^{1,2}(\Omega)$ we have

$$\int_\Omega \nabla u \nabla v dx = \int_D \langle (u_1, u_2), \widetilde{\nabla v} \rangle dx = \lim_{n\to\infty} \int_D \widetilde{\nabla u_n}\widetilde{\nabla v_n} dx = \lim_{n\to\infty} \int_{\Omega_n} g v_n dx = \int_\Omega g v dx. \tag{9}$$

Hence $u|_\Omega = u_{\Omega,g}$ and, moreover,

$$\int_D |\widetilde{\nabla u_{\Omega,g}}|^2 dx = \int_U g u_{\Omega,g} dx \longleftarrow_{n\to\infty} \int_U g u_{\Omega_n,g} dx = \int_D |\widetilde{\nabla u_{\Omega_n,g}}|^2 dx. \tag{10}$$

By the uniqueness of the solution of (1), the whole sequence $\widetilde{\nabla u_{\Omega_n,g}}$ converges to $\widetilde{\nabla u_{\Omega,g}}$ in $L^2(D,\mathbb{R}^2)$.

$2 \Rightarrow 3$. Let C be a connected component of Ω and denote

$$Y = \{\psi \in L^{1,2}(C\backslash\overline{B}) : \exists \psi_n \in L^{1,2}(\Omega_n\backslash\overline{B}) \text{ such that } \widetilde{\nabla\psi_n} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla\psi}\} \subseteq L^{1,2}(C\backslash\overline{B}).$$

Let $\Psi \in L^{1,2}(C \setminus \overline{B})$, $\Psi \perp Y$, i.e., $\int_{C\setminus\overline{B}} \nabla\Psi\nabla v dx = 0$ for all $v \in Y$; let us fix a representative of Ψ in $\mathcal{L}^{1,2}(C \setminus \overline{B})$. Using property (8) of the H^c-convergence we deduce, as above, that $-\Delta\Psi = 0$ in $\mathcal{D}'(C \setminus \overline{B})$. Since every solution $v_{\Omega,\gamma}$ belongs to Y, writing the orthogonality property we get

$$0 = \int_{C\setminus\overline{B}} \nabla\Psi\nabla v_{\Omega,\gamma} dx = \int_{\partial B} \gamma\Psi d\sigma.$$

This relation holds for every $\gamma \in H^{\frac{1}{2}}(\partial B)$ such that $\int_{\partial B}\gamma d\sigma = 0$. Since $H^{\frac{1}{2}}(\partial B)$ is dense in $L^2(\partial B)$ we get that Ψ is constant on ∂B. Now let $\bar\Psi \in L^{1,2}(C)$ such that $\bar\Psi = \Psi$ in $C \setminus \bar B$ and $\bar\Psi = c$ a.e. on B. Since $\Omega_n \overset{H^c}{\longrightarrow} \Omega$, for every function $\varphi \in C_0^\infty(C)$ the restriction $\varphi_{|\Omega\setminus\bar B}$ belongs to Y, and hence we have

$$\int_\Omega \nabla\bar\Psi\nabla\varphi dx = 0.$$

Therefore the extension of Ψ by the same constant on B gives a harmonic function constant on a set of strictly positive measure; hence $\nabla\Psi = 0$ on $\Omega \setminus \overline{B}$. We conclude that Y is dense in $L^{1,2}(C \setminus \bar B)$.

To prove that for all $u \in L^{1,2}(C)$, there exists $u_n \in L^{1,2}(\Omega_n)$ such that $\widetilde{\nabla u_n} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla u}$ we use an argument based on the partition of unity of D. Let $\varphi \in C_0^\infty(C)$ such that $\varphi = 1$ on B. Let $u_n = \tilde u\varphi + (1-\varphi)\tilde v_n$, where $v_n \in L^{1,2}(\Omega\setminus\bar B)$ and $\widetilde{\nabla v_n} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla u_{|C\setminus\bar B}}$. So $u_n \in L^{1,2}(\Omega_n)$ and $\widetilde{\nabla u_n} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla u}$.

Now let $(C_i)_{i\in\mathbb{N}}$ be the family of all connected components of Ω. Since the set

$$\{u \in L^{1,2}(\Omega) \text{ such that } \nabla u = 0 \text{ on } C_i \text{ except for a finite number of } i\}$$

is dense in $L^{1,2}(\Omega)$ assertion 3 follows.

$3 \Rightarrow 2$. The proof follows the same arguments as $3 \Rightarrow 1$ with the remark that every function of $L^{1,2}(\Omega \setminus \overline{B})$ has an extension on $L^{1,2}(\Omega)$. □

3.2. Sufficient topological constraints for the shape stability of (3). In what follows, we use the harmonic conjugates of the solutions of (3) in order to transform the shape continuity problem for (3) into a shape continuity problem of an elliptic equation with Dirichlet boundary conditions. The main reason for doing this is that the study of the domain variation for Dirichlet problems has a complete answer in the case that the Dirichlet boundary condition is zero (or a restriction of some fixed H^1-function). Either necessary and sufficient conditions for shape stability are given

in this case (see [7, 5, 14]) or relaxation results can be established (see [8, 15]). In the latter case, one can describe exactly the "lack" of continuity. Unfortunately, in our case, the Dirichlet boundary condition is not zero, and the values of the functions on different connected components of the boundaries are constants which *vary* with the domains.

That is why we restrict ourselves to the case that for every $n \in \mathbb{N}$ the set $\mathbb{R}^2 \setminus \Omega_n$ has a uniformly bounded number of connected components. In this particular case, we can establish a continuity result for the Dirichlet problem even if the constants are different on each connected component. Nevertheless, some of the results of this section are true without any restriction.

THEOREM 3.2. *If $\{\Omega_n\}_{n\in\mathbb{N}} \in \mathcal{O}_l(D)$ is such that $\Omega_n \xrightarrow{H^c} \Omega$, then for every ball B such that $\overline{B} \subseteq \Omega$ and for every $\gamma \in H^{\frac{1}{2}}(\partial B)$ such that $\int_{\partial B} \gamma d\sigma = 0$, we have*

$$\widetilde{\nabla v_{\Omega_n,\gamma}} \xrightarrow{L^2(D,\mathbb{R}^2)} \widetilde{\nabla v_{\Omega,\gamma}}.$$

Proof. The proof is divided into three steps. We use the duality argument to transform the Neumann problem into a Dirichlet problem, then use continuity results for the domain variation of a Dirichlet problem (such as, for example, a Šverák-type result adapted to different constants), and then return to the Neumann problem.

Step 1. *Passage from the Neumann problem to the Dirichlet problem.* For the existence of the conjugate function into a smooth domain with a finite number of (smooth) holes, we refer to [19, Theorem 3.1]. Since we do not impose any regularity of the boundary (besides the constraint on the number of connected components), we prove in what follows that if on the nonsmooth part of the boundary the normal derivative is zero in the weak sense given by the variational formulation, one can still use a duality argument and modify the result of [19] in order to find a conjugate function with a constant trace on the connected components of the complementary. The sense of the trace on a nonsmooth set is understood as the usual restriction of a quasi-continuous representative of an $H_0^1(D)$-function, this restriction being defined quasi-everywhere (q.e.). (See [20] for details concerning capacity.)

Let $\Omega \in \mathcal{O}_l(D)$ such that $\overline{B} \subseteq \Omega$ and denote by $K_1, \dots, K_l$ the connected components of its complement. Consider (3) on $\Omega \setminus \overline{B}$:

$$\begin{cases} -\Delta v_{\Omega,\gamma} = 0 & \text{in} \quad \Omega \setminus \overline{B}, \\ \frac{\partial v_{\Omega,\gamma}}{\partial n} = 0 & \text{on} \quad \partial\Omega, \\ \frac{\partial v_{\Omega,\gamma}}{\partial n} = \gamma & \text{on} \quad \partial B. \end{cases} \tag{11}$$

If Ω is not connected in every connected component which does not contain B, the solution is set to be 0.

LEMMA 3.3. *There exist a function $\phi \in H_0^1(D)$ and constants $c_1, \dots, c_l \in \mathbb{R}$ such that $\nabla v_{\Omega,g} = \operatorname{curl}\phi$ in $\Omega \setminus \overline{B}$ and*

$$\begin{cases} -\Delta\phi = 0 & \text{in} \quad \Omega \setminus \overline{B}, \\ \phi = c_i & \text{q.e. on} \quad K_i, \quad i = 1, \dots, l, \\ \phi = G & \text{on} \quad \partial B, \end{cases} \tag{12}$$

where $G \in H^{\frac{3}{2}}(\partial B)$ is such that $G' = \gamma$ in the sense of distribution on ∂B.

The equality $\phi = c_i$ q.e. on K_i means that the usual restriction (defined q.e.) of a quasi-continuous representative of ϕ in $H_0^1(D)$ is equal to c_i on K_i.

Proof. We suppose that for every $i = 1, \dots, l$, $diam(K_i) > 0$; if not, we simply ignore K_i since it has zero capacity. If $\partial\Omega$ were Lipschitz, then the result of this lemma would be a straightforward consequence of [19, Theorem 3.1]. Since no assumption on the regularity of the $\partial\Omega$ is made, we consider a sequence of smooth open sets $\{U_n\}_{n\in\mathbb{N}}$ such that $\sharp U_n^c \leq l$, $B(0, r+\delta) \subseteq U_n \subseteq U_{n+1} \subseteq \Omega$, and $\Omega = \bigcup_{n\in\mathbb{N}} U_n$. Let us denote by v_n the solution of the following problem:

$$\left\{\begin{array}{rcl} -\Delta v_n = 0 & \text{in} & U_n \setminus \overline{B}, \\ \frac{\partial v_n}{\partial n} = 0 & \text{on} & \partial U_n, \\ \frac{\partial v_n}{\partial n} = \gamma & \text{on} & \partial B. \end{array}\right. \tag{13}$$

Following [19], there exists a function $\phi_n \in H_0^1(D)$ (obtained by extension with zero on the infinite connected component of $\mathbb{R}^2 \setminus U_n$) such that $\nabla v_n = \mathrm{curl}\phi_n$ in $U_n \setminus \overline{B}$ and

$$\left\{\begin{array}{rcl} -\Delta \phi_n = 0 & \text{in} & U_n \setminus \overline{B}, \\ \phi_n = c_i^n & \text{on} & K_i^n, \quad i = 1, \dots, l, \\ \phi_n = G + c_n & \text{on} & \partial B. \end{array}\right. \tag{14}$$

Taking v_n as a test function in (13) and using the trace theorem and the Poincaré inequality in $H^1(B(0, r+\delta) \setminus \overline{B})$, there exists a constant C independent of n such that

$$\Big(\int_{U_n \setminus B} |\nabla v_n|^2 dx \Big)^{\frac{1}{2}} \leq C |\gamma|_{L^2(\partial B)}.$$

In the connected components of U_n not containing B we have $\nabla v_n = \nabla \phi_n = 0$. Hence $\nabla \phi_n$ is bounded in $L^2(D, \mathbb{R}^2)$. Since ϕ_n is defined up to a constant in each connected component of $U_n \setminus \overline{B}$, we choose the constants such that ϕ_n extended by these constants on each connected component of $\overline{D} \setminus (\Omega \setminus \overline{B})$ belongs to $H_0^1(D)$ (see [20]). These extended functions are denoted by the same symbols. The Poincaré inequality in $H_0^1(D)$ gives that the sequence $\{\phi_n\}_{n\in\mathbb{N}}$ is bounded in $H_0^1(D)$. There exists a function $\phi \in H_0^1(D)$ such that for a subsequence (still denoted with the same index) we can write

$$\nabla \phi_n \overset{L^2(D,\mathbb{R}^2)}{\rightharpoonup} \nabla \phi.$$

As a consequence of the H^c-convergence, property (8), we obtain that $-\Delta\phi = 0$ in $\Omega \setminus \overline{B}$, and by the Banach–Saks theorem we get that $\phi = c_i$ q.e. on K_i, $\phi = G + c$ on ∂B. These equalities hold, since $\{c_i^n\}_{n\in\mathbb{N}}$ is bounded and $cap(K_n^i)$ does not converge to zero (in fact we have $\liminf_{n\to\infty} cap(K_n^i) \geq cap(K) > 0$).

Let us prove that $\nabla v_{\Omega,\gamma} = \mathrm{curl}\phi$ in $\Omega \setminus \overline{B}$. It is sufficient to prove that

$$\widetilde{\nabla v_n} \overset{L^2(D,\mathbb{R}^2)}{\rightharpoonup} \widetilde{\nabla v_{\Omega,\gamma}}.$$

This comes back to proving that the Neumann problem is shape stable to increasing sequences of domains. For a subsequence (still denoted by the same index) we can write

$$\widetilde{\nabla v_n} \overset{L^2(D,\mathbb{R}^2)}{\rightharpoonup} (v_1, v_2).$$

Since U_n is increasing, we get $\partial_2 v_1 = \partial_1 v_2$ in $\Omega \setminus \overline{B}$, and by the De Rham theorem there exists $\bar{v} \in L^2_{loc}(\Omega \setminus \overline{B})$ such that $\nabla \bar{v} = (v_1, v_2)$; hence $\bar{v} \in L^{1,2}(\Omega \setminus \overline{B})$. Moreover, we have that $\bar{v}$ is a weak variational solution of (13) on $\Omega \setminus \overline{B}$ since

$$\begin{aligned}\frac{1}{2}\int_{\Omega\setminus\overline{B}} |\nabla \bar{v}|^2 dx - \int_{\partial B} \gamma \bar{v} d\sigma &\leq \liminf_{n\to\infty} \frac{1}{2}\int_{U_n\setminus\overline{B}} |\nabla v_n|^2 dx - \int_{\partial B} \gamma v_n d\sigma \\ &\leq \liminf_{n\to\infty} \frac{1}{2}\int_{U_n\setminus\overline{B}} |\nabla \xi|^2 dx - \int_{\partial B} \gamma \xi d\sigma \\ &= \frac{1}{2}\int_{\Omega\setminus\overline{B}} |\nabla \xi|^2 dx - \int_{\partial B} \gamma \xi d\sigma,\end{aligned}$$

where $\xi \in L^{1,2}(\Omega \setminus \overline{B})$ is an arbitrary element. Consequently, we get that $\bar{v} = v_{\Omega,\gamma}$. □

Step 2. *Continuity with respect to the domain variation for the associated Dirichlet problems.* We give without proofs two technical lemmas. The first one is an immediate consequence of [4, 7], while the second one can be proved using circular rearrangements (see [13]) and noticing that in one dimension the step functions are not in $H^{\frac{1}{2}}(\mathbb{R})$.

LEMMA 3.4. *Let $\{\phi_n\}_{n\in\mathbb{N}} \subseteq H^1_0(D)$, let $\{K_n\}_{n\in\mathbb{N}}$ be a sequence of compact connected sets in D, and let $\{c_n\}_{n\in\mathbb{N}}$ be a sequence of constants such that $\phi_n(x) = c_n$ q.e. on K_n. If*

$$K_n \xrightarrow{H} K \qquad \text{and} \qquad \phi_n \overset{H^1_0(D)}{\rightharpoonup} \phi,$$

there exists a constant $c \in \mathbb{R}$ such that $c_n \longrightarrow c$ and $\phi(x) = c$ q.e. on K.

LEMMA 3.5. *Let $\phi \in H^1_0(D)$ and let K_1, K_2 be two compact connected sets in D with positive diameter. If there exist two constants $c_1, c_2 \in \mathbb{R}$ such that $\phi(x) = c_1$ q.e. on K_1 and $\phi(x) = c_2$ q.e. on K_2, then $K_1 \cap K_2 = \emptyset$.*

Let us assume that $\{\Omega_n\}_{n\in\mathbb{N}}$ is a sequence satisfying the hypotheses of Theorem 3.2. As in the previous step, we denote by ϕ_n, ϕ the corresponding functions found by Lemma 3.3 applied to $v_{\Omega_n,\gamma}$ on Ω_n and $v_{\Omega,\gamma}$ on Ω, respectively. We denote the connected components of $\overline{D} \setminus \Omega_n$ by $K^n_1, \ldots, K^n_l$, some of them perhaps being empty.

LEMMA 3.6. *There exist a subsequence $\{\phi_{n_k}\}_{k\in\mathbb{N}}$ such that*

$$\phi_{n_k} \overset{H^1_0(D)}{\rightharpoonup} \phi$$

and a function $v \in L^{1,2}(\Omega \setminus \overline{B})$ such that $\operatorname{curl}\phi = \nabla v$ *in* $\Omega \setminus \overline{B}$.

Proof. Since the extension by constants of ϕ_n does not increase the norm of the gradient, and since we have $\int_{\Omega_n\setminus\overline{B}} |\nabla \phi_n|^2 dx = \int_{\Omega_n\setminus\overline{B}} |\nabla u_n|^2 dx$, we get that $\{\widetilde{\nabla\phi_n}\}_{n\in\mathbb{N}}$ is bounded in $L^2(D, \mathbb{R}^2)$. Hence for a subsequence we have

$$\phi_{n_k} \overset{H^1_0(D)}{\rightharpoonup} \phi.$$

From the Hausdorff convergence we get $-\Delta\phi = 0$ in $\Omega \setminus \overline{B}$.

Without losing the generality, we can suppose that for a subsequence (still denoted by the same index) and for all $i = 1, \ldots, l$ we have $K^{n_k}_i \xrightarrow{H} K_i$. Using Lemma 3.4 we also get $c_{n_k,i} \to c_i$ and $\phi = c_i$ q.e. on K_i. If there exist two compact sets with positive diameter K_{i_1} and K_{i_2} and nonempty intersection, then from Lemma 3.5 we get that $c_{i_1} = c_{i_2}$.

Since $\overline{D}\setminus\Omega = \cup_{i=1}^{l} K_i$ we get that ϕ is constant q.e. on every connected component of $\overline{D}\setminus\Omega$. Using property (7) of the H^c-convergence, there exists $v \in L^{1,2}(\Omega)$ such that

$$\widetilde{\nabla v_{\Omega_n,\gamma}} \overset{L^2(D,\mathbb{R}^2)}{\rightharpoonup} (v_1, v_2)$$

and $\nabla v = (v_1, v_2)$ in Ω. The relation $\nabla v_{\Omega_n,\gamma} = \mathrm{curl}\phi_n$ in Ω_n gives (again from property (7)) that $\nabla v = \mathrm{curl}\phi$ in $\Omega \setminus \overline{B}$. □

Step 3. *Passage from the Dirichlet problem to Neumann problem under the rotational hypothesis.* The result obtained in the previous step asserts that the weak limit ϕ is such that $-\Delta\phi = 0$ in $\Omega\setminus\overline{B}$ and ϕ is q.e. constant on each connected component of $\overline{D}\setminus\Omega$. In what follows we prove that ϕ has a conjugate; i.e., ϕ is exactly the function obtained by applying Lemma 3.3 to $v_{\Omega,\gamma}$ on $\Omega\setminus\overline{B}$.

LEMMA 3.7. *Let O be a smooth open connected set and K a compact connected set such that $K \subseteq O$. Let us denote by θ the capacitary potential of K in O, i.e., the function $\theta \in H_0^1(O)$ such that*

$$\begin{cases} -\Delta\theta = 0 & \text{in} \quad O\setminus K, \\ \theta = 0 & \text{on} \quad \partial O, \\ \theta = 1 & \text{q.e. on} \quad \partial K. \end{cases} \tag{15}$$

There does not exist a function $\xi \in L^{1,2}(O\setminus K)$ such that $\mathrm{curl}\theta = \nabla\xi$, *unless* $diam(K) = 0$.

Proof. Suppose $diam(K) > 0$. Since O is smooth and θ attains its minimum in all the points of ∂O, we get by the Hopf maximum principle that $\frac{\partial\theta}{\partial n} \neq 0$ in any point of ∂O. There exists a neighborhood of ∂O of the form $\{\theta \le c\}$ with $c > 0$ in which $|\nabla\theta| \neq 0$. Indeed, supposing that there exists x_n such that $\theta(x_n) = c_n \to 0$ and $\nabla\theta(x_n) = 0$, we have by compactness that for a subsequence (still denoted with the same index) $x_n \to x$; therefore, $\theta(x) = 0$, and hence $x \in \partial O$. On the other side, the gradient is also continuous up to the boundary, which yields $\nabla\theta(x) = 0$, contrary to the previous assertion.

Let us denote $U = \{x \in O : \theta(x) < c\}$. The open set U has a smooth boundary, $\partial O \cup \{\theta = c\}$. Computing

$$\int_U |\nabla\theta|^2 dx = \int_{\partial O\cup\{\theta=c\}} \theta\frac{\partial\theta}{\partial n} d\sigma - \int_U \theta\Delta\theta dx = \int_{\{\theta=c\}} \theta\frac{\partial\theta}{\partial n} d\sigma.$$

Using the hypothesis $\mathrm{curl}\theta = \nabla\xi$, we can write

$$\int_{\{\theta=c\}} \theta\frac{\partial\theta}{\partial n} d\sigma = \int_{\{\theta=c\}} c\frac{\partial\xi}{\partial t} d\sigma = 0.$$

Here $\frac{\partial\xi}{\partial t}$ denotes the tangential derivative of ξ to the smooth curve $\{\theta = c\}$. Therefore, we would get that θ vanishes on U, contrary to the maximum principle. □

LEMMA 3.8. *Let $\Omega \in \mathcal{O}_l(D)$ such that $\overline{B} \subseteq \Omega$. Suppose that there exist a function $\phi \in H_0^1(D)$ and a function $u \in L^{1,2}(\Omega\setminus\overline{B})$ such that $\nabla u = \mathrm{curl}\phi$ in $\Omega\setminus\overline{B}$ and*

$$\begin{cases} -\Delta\phi = 0 & \text{in} \quad \Omega\setminus\overline{B}, \\ \phi = c_i & \text{on} \quad K_i, \quad i = 1,\dots,l, \\ \phi = G + c & \text{on} \quad \partial B. \end{cases} \tag{16}$$

Then u is the weak solution of (11) *on $\Omega\setminus\overline{B}$.*

Proof. Since $u \in L^{1,2}(\Omega \setminus \overline{B})$ it suffices to prove that for any $\xi \in L^{1,2}(\Omega \setminus \overline{B})$ we have

$$\int_{\Omega\setminus\overline{B}} \nabla u \nabla \xi dx = \int_{\partial B} \gamma \xi d\sigma.$$

Considering smooth neighborhoods O_i of K_i, by an argument of partition of unity, it suffices to prove that for any function $\xi \in H^1(O_i \setminus K_i)$ vanishing q.e. on ∂O_i we have

$$\int_{O_i\setminus K_i} \nabla u \nabla \xi dx = 0.$$

It suffices actually to prove that u is a solution of the following problem on $O_i \setminus K_i$:

$$\left\{ \begin{array}{rcl} -\Delta u = 0 & \text{in} & O_i \setminus K_i, \\ \frac{\partial u}{\partial n} = 0 & \text{on} & \partial K_i, \\ \frac{\partial u}{\partial n} = \frac{\partial \phi}{\partial t} & \text{on} & \partial O_i. \end{array} \right. \tag{17}$$

To the solution u^* of this equation we associate the function ϕ^* given by Lemma 3.3. We have that $-\Delta\phi^* = 0$ in $O_i \setminus K_i$, $\phi^* = \phi$ on ∂O_i, $\phi^* = c^*$ on K_i. Denoting $\theta = \phi - \phi^*$, we get that $\nabla\theta = \text{curl}(u - u^*)$, $-\Delta\theta = 0$ in $O_i \setminus K_i$, $\theta = 0$ on ∂O_i, $\theta = c - c^*$ on K_i. Following Lemma 3.7, since $diam(K_i) > 0$, we get $c = c^*$, and hence $u = u^*$. □

Remark that Step 1 is in general true, without any assumption on the number of connected components. Indeed, when taking the approximating sequence $\{u_n\}_{n\in\mathbb{N}}$ of smooth functions, on each set U_n, [19, Theorem 3.1] can be applied. Step 2 fails to be true in general.

4. The general case $a \geq 0$. The following result establishes the relation between the shape stability of problems (1) and (3). This result is general and does not require any geometrical or topological constraints on Ω_n.

THEOREM 4.1. *Given a sequence of open sets $\{\Omega_n\}_{n\in\mathbb{N}}$ converging in the Hausdorff complementary topology to Ω, assertions* (A) *and* (B) *are equivalent:*

(A) *For every admissible h, we have*

$$(\tilde{u}_{\Omega_n,h}, \widetilde{\nabla u_{\Omega_n,h}}) \overset{L^2_a(D)\times L^2(D,\mathbb{R}^2)}{\longrightarrow} (\tilde{u}_{\Omega,h}, \widetilde{\nabla u_{\Omega,h}}).$$

(B) *The following three conditions hold:*

(B.1) *For every ball B such that $\overline{B} \subseteq \Omega$ and for every $\gamma \in H^{\frac{1}{2}}(\partial B)$ with $\int_{\partial B} \gamma d\sigma = 0$, we have*

$$\widetilde{\nabla v_{\Omega_n,\gamma}} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla v_{\Omega,\gamma}}.$$

(B.2) *For every $u \in L^{1,2}_a(\Omega)$ such that $\nabla u = 0$, there exist*

$$u_n \in L^{1,2}_a(\Omega_n) \quad \text{such that} \quad (\tilde{u}_n, \widetilde{\nabla u_n}) \overset{L^2_a(D)\times L^2(D,\mathbb{R}^2)}{\longrightarrow} (\tilde{u}, 0).$$

(B.3) $|\Omega \cap \{a > 0\}| = \lim_{n\to\infty} |\Omega_n \cap \{a > 0\}|$.

When investigating the shape stability of (1), condition (B.1) is, in general, the difficult one. As seen in section 3, the duality argument can be applied successfully

in some particular situations. Condition (B.2) is easy to handle as soon as Ω is connected, and condition (B.3) is trivial to check in concrete examples.

Proof. (A) $\Rightarrow$ (B) For proving (B.1) it is enough to prove assertion 3 of Proposition 3.1. Take $u \in L^{1,2}(\Omega)$ and define for every $M > 0$

$$u_M := \min\{\max\{u^*, -M\}, M\},$$

where u^* is a representative of u in $\mathcal{L}^{1,2}(\Omega)$. Then u_M converges in $L^{1,2}(\Omega)$ to u when $M \to +\infty$ and, moreover, u_M belongs to $L^{1,2}_a(\Omega)$.

Let us denote

$$Y = \{u_{\Omega,h} : h \text{ admissible}\} \subseteq L^{1,2}_a(\Omega)$$

and prove that this set is dense in $L^{1,2}_a(\Omega)$. Suppose for contradiction that $u \in L^{1,2}_a(\Omega)$ is orthogonal on Y, i.e.,

$$\int_\Omega \nabla u \nabla u_{\Omega,h} + a u u_{\Omega,h} dx = 0.$$

Consequently

$$\int_\Omega u(af + g)dx = 0.$$

Taking $f = 0$, we get that u is constant on every connected component of Ω, and taking $g = 0$ we get that $au = 0$; hence $\int_\Omega |\nabla u|^2 + au^2 dx = 0$, i.e., $u \equiv 0$ in $L^{1,2}_a(\Omega)$. So, for every $M > 0$ and for every $\varepsilon > 0$ there exists $h_{M,\varepsilon}$ such that

$$\int_\Omega |\nabla u_{\Omega,h_{M,\varepsilon}} - \nabla u_M|^2 dx + \int_\Omega (u_{\Omega,h_{M,\varepsilon}} - u_M)^2 a(x) dx \le \varepsilon.$$

Since from hypothesis (A)

$$\int_D |\widetilde{\nabla u_{\Omega,h_{M,\varepsilon}}} - \widetilde{\nabla u_{\Omega_n,h_{M,\varepsilon}}}|^2 dx + \int_D (\tilde{u}_{\Omega,h_{M,\varepsilon}} - \tilde{u}_{\Omega_n,h_{M,\varepsilon}})^2 a(x) dx \to 0,$$

by a usual diagonal procedure we find a sequence of the form $\{(\tilde{u}_{\Omega_n,h_{M,\varepsilon_n}}, \widetilde{\nabla u_{\Omega_n,h_{M,\varepsilon_n}}})\}$ which converges in $L^2_a(D) \times L^2(D,\mathbb{R}^2)$ to $(\tilde{u}_M, \widetilde{\nabla u_M})$. We finish the proof by taking $M \to \infty$ and observing that for every open set U the injection $L^{1,2}_a(U) \hookrightarrow L^{1,2}(U)$ is a contraction.

To prove (B.2) let us consider $u \in L^{1,2}_a(\Omega)$ such that $\nabla u = 0$. Take $g = 0$ and $f = u$. Then $u = u_{\Omega,h}$, and hypothesis (A) gives the conclusion.

In order to prove (B.3) take $g = 0$ and $f = 1$. Then $u_{\Omega_n,h} = 1_{\Omega_n}$, and hypothesis (A) gives

$$\int_{\Omega_n} a dx \to \int_\Omega a dx. \tag{18}$$

From the H^c-convergence we have $1_\Omega \le \liminf_{n\to\infty} 1_{\Omega_n}$ a.e. in D, and hence

$$\liminf_{n\to\infty} 1_{\Omega_n \cap \{a>0\}} \ge 1_{\Omega \cap \{a>0\}}. \tag{19}$$

From (18) and (19) we get $1_{\Omega_n \cap \{a>0\}} \overset{L^1(D)}{\longrightarrow} 1_{\Omega \cap \{a>0\}}$; therefore (B.3) holds.

(B) $\Rightarrow$ (A) Assume in a first step that the set

$$Y = \{\phi \in L^{1,2}_a(\Omega) : \exists \phi_n \in L^{1,2}_a(\Omega_n) \text{ such that } (\tilde{\phi}_n, \widetilde{\nabla \phi_n}) \to (\tilde{\phi}, \widetilde{\nabla \phi}) \quad L^2_a(D) \times L^2(D, \mathbb{R}^2) - \text{strongly}\} \tag{20}$$

is dense in $L^{1,2}_a(\Omega)$. Then (A) follows straightforwardly. Indeed, by the boundedness of $\{(\tilde{u}_{\Omega_n,h}, \widetilde{\nabla u_{\Omega_n,h}})\}$ in $L^2_a(D) \times L^2(D, \mathbb{R}^2)$ there exists a subsequence (still denoted by the same index) such that

$$(\tilde{u}_{\Omega_n,h}, \widetilde{\nabla u_{\Omega_n,h}}) \rightharpoonup (u, u_1, u_2) \quad \text{in} \quad L^2_a(D) \times L^2(D, \mathbb{R}^2) - \text{weakly}. \tag{21}$$

From property (8) of the Hausdorff convergence and the De Rham theorem, we get that $\nabla u = (v_1, v_2)$ on Ω. Let us fix $\phi \in Y$. Writing the fact that $u_{\Omega_n,h}$ is a solution on Ω_n with ϕ_n as a test function (where ϕ_n is given by (20)) and passing to the limit for $n \to \infty$, by the usual pairing (weak, strong)-convergence we get

$$\int_D \langle (u_1, u_2), \widetilde{\nabla \phi} \rangle dx + \int_D u \tilde{\phi} a dx = \int_D h \tilde{\phi} dx.$$

Since $\widetilde{\nabla \phi}(x) = 0$ on $D \setminus \Omega$ and $\tilde{\phi} = 0$ on $D \setminus \Omega$, we have

$$\int_\Omega \nabla u \nabla \phi dx + \int_\Omega u \phi dx = \int_\Omega h \phi dx. \tag{22}$$

Because (22) holds for every $\phi \in Y$ and Y is dense in $L^{1,2}_a(\Omega)$, we get that $u|_\Omega = u_{\Omega,h}$ and $(u_1, u_2)|_\Omega = \nabla u_{\Omega,h}$. Taking $u_{\Omega_n,h}$ as a test function on Ω_n and passing to the limit for $n \to \infty$, we have

$$\int_{\Omega_n} |\nabla u_{\Omega_n,h}|^2 + a u^2_{\Omega_n,h} dx = \int_{\Omega_n} h u_{\Omega_n,h} dx = \int_D h \tilde{u}_{\Omega_n,h} dx \to \int_D h u dx. \tag{23}$$

We wrote in the previous equality $\int_D h u dx = \int_\Omega h u dx$. Indeed, $\int_D g u dx = \int_\Omega g u dx$ because supp $g \subseteq \Omega$, and $\int_D f u a dx = \int_\Omega f u a dx$ because $au = 0$ on $\Omega^c \cap \{a > 0\}$. The last inequality is a direct consequence of hypothesis (B.3). Since $u = u_{\Omega,h}$ on Ω, using relation (23) we get

$$|(u, u_1, u_2)|_{L^2_a(D) \times L^2(D,\mathbb{R}^2)} \geq |(\tilde{u}_{\Omega,h}, \widetilde{\nabla u_{\Omega,h}})|_{L^2_a(D) \times L^2(D,\mathbb{R}^2)}$$

$$= \lim_{n \to \infty} |(\tilde{u}_{\Omega_n,h}, \widetilde{\nabla u_{\Omega_n,h}})|_{L^2_a(D) \times L^2(D,\mathbb{R}^2)} \geq |(u, u_1, u_2)|_{L^2_a(D) \times L^2(D,\mathbb{R}^2)}.$$

Consequently, we get that $(u, u_1, u_2) = (\tilde{u}_{\Omega,h}, \widetilde{\nabla u_{\Omega,h}})$ and that convergence (21) is strong. The continuity $u_{\Omega_n,h} \to u_{\Omega,h}$ was proved for a subsequence, but since the limit is unique, it holds for the whole sequence.

To finish, let us prove that Y is dense in $L^{1,2}_a(\Omega)$. By linearity and a truncation argument, we can fix $\phi \in L^{1,2}_a(\Omega)$ such that $\phi \in L^\infty(\Omega)$ and $\phi = 0$ on $\Omega \setminus C$, where C is one connected component of Ω. From now on, we fix one representative of ϕ in $\mathcal{L}^{1,2}_a(\Omega)$. Following (B.1) there exists $u_n \in L^{1,2}(\Omega_n)$ such that

$$\widetilde{\nabla u_n} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla \phi}.$$

Let us fix a ball B such that $\overline{B} \subseteq C$, and choosing the representative of u_n in $\mathcal{L}^{1,2}(\Omega_n)$ by adding a suitable constant, we can assume that $\int_B u_n dx = \int_B \phi dx$. Let M be a positive constant such that $\|\phi\|_\infty < M$ and define

$$u_n^M = \max\{\min\{u_n, M\}, -M\}.$$

We notice that $u_n^M \in L_a^{1,2}(\Omega_n)$ and

$$\widetilde{\nabla u_n^M} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla \phi}.$$

Moreover, since $\{\tilde{u}_n^M\}_n$ is uniformly bounded in $L^\infty(D)$, we can write (for a subsequence)

$$\tilde{u}_{n_k}^M \overset{L^2(D)}{\longrightarrow} v,$$

where $\nabla v = \nabla \phi$ on Ω and $v = \phi$ on C. Using the Poincaré inequality on smooth open subsets compactly contained in C, we have that the convergence is strong in $L^2_{loc}(C)$.

Following (B.2), there exists $v_{n_k} \in L_a^{1,2}(\Omega_{n_k})$ such that

$$(\tilde{v}_{n_k}, \widetilde{\nabla v_{n_k}}) \overset{L_a^2(D)\times L^2(D,\mathbb{R}^2)}{\longrightarrow} (\tilde{v}_{|\Omega} - \tilde{\phi}, 0).$$

It is obvious that v_{n_k} can be chosen such that $\|v_{n_k}\|_\infty \leq 2M$. Let us define $\phi_{n_k} := u_{n_k}^M - v_{n_k} \in L_a^{1,2}(\Omega_{n_k})$. We have

$$\widetilde{\nabla \phi_{n_k}} \overset{L^2(D,\mathbb{R}^2)}{\longrightarrow} \widetilde{\nabla \phi}.$$

Let us prove that $\int_D a(\tilde{\phi}_{n_k} - \tilde{\phi})^2 dx \to 0$. First, we have

$$\int_\Omega a(\tilde{\phi}_{n_k} - \tilde{\phi})^2 dx \to 0$$

since on every compact set $\omega \subseteq \Omega$ we have that $\tilde{\phi}_{n_k} - \tilde{\phi}$ weakly converges to 0 in $L^2(\omega)$, the gradient converges to zero strongly, and the sequence is uniformly bounded in $L^\infty(D)$. Second,

$$\int_{D\setminus\Omega} a(\tilde{\phi}_{n_k} - \tilde{\phi})^2 dx \leq 4M^2 \int_{D\setminus\Omega} a 1_{\Omega_{n_k}} dx,$$

the last term converging to zero from (B.3).

Notice that we found a subsequence $\{\phi_{n_k}\}$ and not a sequence converging to ϕ. Suppose for contradiction that there does not exist a sequence $\{\phi_n\}$ strongly converging to ϕ. For a subsequence, we would have that the distance in $L_a^2(D) \times L^2(D,\mathbb{R}^2)$ from ϕ to $L_a^{1,2}(\Omega_{n_k})$ would be bounded below by a positive number. This is absurd since, using the same arguments, there exists a subsequence of this subsequence for which the property cannot hold. □

An immediate consequence of Theorems 4.1 and 3.2 is Theorem 1.1 announced in the introduction.

Proof of Theorem 1.1.

Necessity. Following Theorem 4.1, condition (B.3) holds.

Sufficiency. Let us prove that (B.1), (B.2), and (B.3) hold. Condition (B.1) is a consequence of Theorem 3.2, and condition (B.3) is assumed by hypothesis. One has

only to verify condition (B.2) of Theorem 4.1. If Ω is connected, this is trivial, since every function with zero gradient in Ω is constant, say $c1_\Omega$. Therefore, the sequence $c1_{\Omega_n}$ solves (B.2) using the hypothesis on the convergence of the measures in the region where a is positive. If Ω is not connected, then condition (B.2) is a consequence of the more involved geometric argument relating the Hausdorff convergence to the capacity. We recall this result from [6].

LEMMA 4.2. *If $\{\Omega_n\}_{n\in\mathbb{N}}$ is a sequence of simply connected open sets in D such that $\Omega_n \xrightarrow{H^c} \Omega_a \cup \Omega_b$, where $\Omega_a \cap \Omega_b = \emptyset$, then there exist a subsequence (still denoted by the same index) and two sequences of simply connected open sets $\{\Omega_n^a\}_{n\in\mathbb{N}}$, $\{\Omega_n^b\}_{n\in\mathbb{N}}$ such that $\Omega_n^a \cap \Omega_n^b = \emptyset$, $\Omega_n^a \cup \Omega_n^b \subseteq \Omega_n$, $cap(\Omega_n \setminus (\Omega_n^a \cup \Omega_n^b)) \to 0$, and $\Omega_n^a \xrightarrow{H^c} \Omega_a$, $\Omega_n^b \xrightarrow{H^c} \Omega_b$.*

Using this lemma, condition (B.2) can be proved using a partition of the unity and localizing around the boundary of $\partial\Omega$, as in [9]. □

5. Further remarks. In what follows we give a simple example showing the influence of the positivity of the function a on the shape stability of (1).

Example 5.1. Let

$$\Omega_n = (-1,1)\times(0,1) \setminus (-1,0] \times \left\{ \frac{k}{n} : k = 1, \dots, n-1 \right\}.$$

Note that $\Omega_n \xrightarrow{H^c} \Omega := (0,1)\times(0,1)$.

Take $a_1 = 1_{[0,1]^2}$. Then, following Theorem 1.1, shape stability holds for every admissible h. For example, if $h = 1_{[0,1]^2}$, then $u_{\Omega_n,h} = 1_{\Omega_n}$. Clearly, $u_{\Omega_n,h}$ converges in $L^2_{a_1}(D) \times L^2(D,\mathbb{R}^2)$ to $u_{\Omega,h}$. This can be directly verified, since $\widetilde{\nabla 1_{\Omega_n}} \equiv 0$ and $1_{\Omega_n}|_{\{a_1>0\}} \to 1_\Omega|_{\{a_1>0\}}$ strongly in $L^2(\{a_1 > 0\})$.

If $a_2 = 1_{[-1,1]\times[0,1]}$, then, for the same h,

$$u_{\Omega_n,h}(x,y) = \begin{cases} \frac{e^2}{2(1+e^2)}e^x + \frac{1}{2(1+e^2)}e^{-x} & \text{if } (x,y) \in \Omega_n \cap (-1,0]\times(0,1), \\ -\frac{1}{2(1+e^2)}e^x - \frac{e^2}{2(1+e^2)}e^{-x} + 1 & \text{if } (x,y) \in \Omega_n \cap [0,1)\times(0,1), \end{cases}$$

while $u_{\Omega,h} = 1_\Omega$. Clearly, $u_{\Omega_n,h}$ does not converge to $u_{\Omega,h}$.

Remark 5.2 (shape stability and Mosco convergence). In general, when investigating the shape continuity of the solution of some variational PDE, one has to refer to the Mosco convergence of the associated functional spaces. A general result relating the Mosco convergence of functional spaces and the convergence of minima of some functionals is contained in [1, Theorem 3.6.6]. We briefly recall the definitions of the Kuratowski limits and Mosco convergence.

Let X be a Hilbert space and let $\{G_n\}_{n\in\mathbb{N}}$ be a sequence of subsets of X. The weak upper and the strong lower limits in the sense of Kuratowski are defined as follows:

$$w - \limsup_{n\to\infty} G_n = \left\{ u \in X : \exists \{n_k\}_k, \exists u_{n_k} \in G_{n_k} \text{ such that } u_{n_k} \overset{w-X}{\rightharpoonup} u \right\},$$

$$s - \liminf_{n\to\infty} G_n = \left\{ u \in X : \exists u_n \in G_n \text{ such that } u_n \xrightarrow{s-X} u \right\}.$$

If $\{G_n\}_{n\in\mathbb{N}}$ are closed subspaces in X, it is said that G_n converges in the sense of Mosco to G if

(M_1) $G \subseteq s - \liminf_{n\to\infty} G_n$,
(M_2) $w - \limsup_{n\to\infty} G_n \subseteq G$.
Note that in general $s - \liminf_{n\to\infty} G_n \subseteq w - \limsup_{n\to\infty} G_n$. Therefore, if G_n converges in the sense of Mosco to G, then

$$s - \liminf_{n\to\infty} G_n = G = w - \limsup_{n\to\infty} G_n.$$

For our purpose, the space X is, following the embedding given by relations (5)–(6), $L^2_a(D) \times L^2(D, \mathbb{R}^2)$.

It can be easily proved that if $L^{1,2}_a(\Omega_n)$ converges in the sense of Mosco to $L^{1,2}_a(\Omega)$, then for every admissible h, $u_{\Omega_n,h}$ converges to $u_{\Omega,h}$. Moreover, if $a(x) > 0$ a.e. in D, it can be proved that if $\Omega_n \overset{H^c}{\to} \Omega$, then the Mosco convergence is *equivalent* to the shape stability of the solution for every admissible h; this was proved in [6] for $a \equiv 1$. We notice that if a vanishes on some regions, this equivalence fails. We may have shape stability without Mosco convergence. Indeed, in the example above it is enough to take, with $a_1 = 1_{[0,1]^2}$, the sequence of functions $u_n(x, y) = x$. The second Mosco condition is not satisfied, since the weak limit in $L^2_{a_1}(D) \times L^2(D, \mathbb{R}^2)$ of this sequence has a nonvanishing gradient on $D \setminus \Omega$. Although the second Mosco condition is not satisfied in general, we note that it is satisfied for every sequence of solutions.

The first Mosco condition, i.e., every function $u \in L^{1,2}_a(\Omega)$ is a strong limit (in the sense of extensions) of a sequence of functions of $L^{1,2}_a(\Omega_n)$, is implicitly present in Proposition 3.1, condition 3 and in the proof of Theorem 4.1. In concrete situations, this condition is the one that is difficult to handle. If Ω would have a smooth boundary, then the restrictions to Ω_n of any extension of u would straightforwardly give (M_1). However, in general we deal with nonsmooth sets and u might not possess an extension in $L^{1,2}(D)$. If Ω has a crack, the "traces" of the function may be different on each side of the crack.

Remark 5.3. Theorem 1.1 remains valid if the operator $-\Delta$ in (1) is replaced by a general operator A in the divergence form

$$Au = -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\Big(a_{ij}(x)\frac{\partial}{\partial x_j}u\Big),$$

with $(a_{ij}) \in L^\infty(D, \mathbb{R}^4)$, uniformly elliptic.

Of course, the duality argument is no longer valid for this operator; for proving Theorem 1.1 one can use assertion 3 of Proposition 3.1.

For other approaches of the shape stability of (1), through homogenization or relaxation techniques, we refer the reader to [3, 16, 17, 21, 22] and [11], respectively. The notion of weak connected domains of [21], even if it does not appear explicitly here, is strongly related through Lemma 4.2 to the convergence in the sense of Kuratowski of the families of locally constant functions. The general relaxed form of (1) is not known.

Largely studied in the literature is also the continuity with respect to the domain variation of the solution of an elliptic problem with homogeneous Dirichlet boundary conditions (in (1) the Neumann condition is replaced by $u = 0$ on $\partial\Omega$). The complete relaxation result for this problem was obtained in [15]. In a certain way, the study of the shape continuity for Dirichlet problems is easier, mainly because the H^1_0-Sobolev spaces enjoy a very natural extension property, but also because many results of potential theory relating the oscillations of harmonic functions on the boundary to the Wiener criterion can be applied.

Another interesting question is to find whether the spectrum of the Neumann–Laplacian is stable for perturbations of the geometric domain. As shown in the classical example of Courant and Hilbert [12], the spectrum is not stable in the case when a fixed square is perturbed by a small square connected by a channel. Consequently, the resolvent operators do not converge in the operator norm topology, but following Theorem 1.1 they converge strongly.

REFERENCES

[1] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman, Boston, 1984.
[2] H. Brezis, *Analyse Fonctionelle*, Masson, Paris, 1983.
[3] M. Briane, *Homogenization of the torsion problem and the Neumann problem in non regular periodically perforated domains*, Math. Models Methods Appl. Sci., 7 (1997), pp. 847–870.
[4] D. Bucur, *Shape analysis for non smooth elliptic operators*, Appl. Math. Lett., 9 (1996), pp. 11–16.
[5] D. Bucur, *Characterization for the Kuratowski limits of a sequence of Sobolev spaces*, J. Differential Equations, 151 (1999), pp. 1–19.
[6] D. Bucur and N. Varchon, *Boundary variation for the Neumann problem*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 24 (2000), pp. 807–821.
[7] D. Bucur and J.P. Zolésio, *N-dimensional shape optimization under capacitary constraints*, J. Differential Equations, 123 (1995), pp. 504–522.
[8] G. Buttazzo and G. Dal Maso, *Shape optimization for Dirichlet problems: Relaxed formulation and optimality conditions*, Appl. Math. Optim., 23 (1991), pp. 17–49.
[9] A. Chambolle and F. Doveri, *Continuity of Neumann linear elliptic problems on varying two-dimensional bounded open sets*, Comm. Partial Differential Equations, 22 (1997), pp. 811–840.
[10] D. Chenais, *On the existence of a solution in a domain identification problem*, J. Math. Anal. Appl., 52 (1975), pp. 189–289.
[11] G. Cortesani, *Asymptotic behaviour of a sequence of Neumann problems*, Comm. Partial Differential Equations, 22 (1997), pp. 1691–1729.
[12] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 1, Interscience, New York, 1953.
[13] S. Cox and B. Kawohl, *Circular symmetrization and extremal Robin conditions,* Z. Angew. Math. Phys., 50 (1999), pp. 301–311.
[14] G. Dal Maso, *Some necessary and sufficient conditions for the convergence of unilateral convex sets*, J. Funct. Anal., 62 (1985), pp. 119–159.
[15] G. Dal Maso and U. Mosco, *Wiener's criterion and Γ-convergence*, Appl. Math. Optim., 15 (1987), pp. 15–63.
[16] A. Damlamian, *Le problème de la passoire de Neumann*, Rend. Sem. Mat. Univ. Politec. Torino, 43 (1985), pp. 427–450.
[17] T. Del Vecchio, *The thick Neumann's sieve*, Ann. Mat. Pura Appl. (4), 147 (1987), pp. 363–402.
[18] L.C. Evans and R.F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Ann Harbor, MI, 1992.
[19] V. Girault and P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[20] J. Heinonen, T. Kilpelainen, and O. Martio, *Nonlinear potential theory of degenerate elliptic equations*, Clarendon Press, Oxford, New York, Tokyo, 1993.
[21] E.Ya. Khruslov, *Homogenized models of composite media*, in Composite Media and Homogenization Theory, Progr. Nonlinear Differential Equations Appl. 5, Birkhäuser Boston, Boston, 1991, pp. 159–182.
[22] F. Murat, *The Neumann sieve*, in Nonlinear Variational Problems, A. Marino et al., eds., Res. Notes in Math. 127, Pitman, Boston, 1985, pp. 24–32.
[23] P. Neittaanmäki and D. Tiba, *Shape optimization in free boundary systems*, in Free Boundary Problems: Theory and Applications, II (Chiba, 1999), GAKUTO Internat. Ser. Math. Sci. Appl. 14, Gakkōtosho, Tokyo, 2000, pp. 334–343.
[24] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, Berlin, 1984.
[25] V. Šverák, *On optimal shape design*, J. Math. Pures Appl. (9), 72 (1993), pp. 537–551.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 478–494

NONLINEAR STABILITY IN L^p FOR A CONFINED SYSTEM OF CHARGED PARTICLES*

MARÍA J. CÁCERES†, JOSÉ A. CARRILLO†, AND JEAN DOLBEAULT‡

Abstract. We prove the nonlinear stability in L^p, with $1 \le p \le 2$, of particular steady solutions of the Vlasov–Poisson system for charged particles in the whole space $\mathbb{R}^6$. Our main tool is a functional associated to the relative entropy or Casimir-energy functional.

Key words. kinetic equations, Vlasov–Poisson system, nonlinear stability, relative entropy, Csiszár–Kullback inequality, interpolation inequalities

AMS subject classifications. Primary, 35B35, 82D10; Secondary, 35B45, 35D05, 82C40, 82D37, 76X05

PII. S0036141001398435

1. Introduction.

We consider a gas of charged particles described by a distribution function $f(t,x,v) \ge 0$ which represents the probability density of particles at position x with velocity v at time t. The evolution of f is governed by the Liouville evolution equation

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f + F(t,x) \cdot \nabla_v f = 0 \tag{1.1}$$

in $\mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3$, where the electric field $F(t,x)$ is given by an external potential ϕ_e and by a mean field potential ϕ according to

$$F(t,x) \;=\; -q\left(\nabla_x \phi(t,x) + \nabla_x \phi_e(x)\right) . \tag{1.2}$$

The electrostatic potential $\phi \ge 0$ is self-consistently computed by

$$\phi = K * \rho(f) \tag{1.3}$$

with $K = \frac{q}{4\pi\epsilon_0}|x|^{-1}$, where $\rho(f)$ is the spatial density of particles, which is defined by

$$\rho(f)(t,x) \;=\; \int_{\mathbb{R}^3} f(t,x,v)\; dv\ .$$

As usual, ϵ_0 and q are, respectively, the permittivity of the vacuum and the elementary charge of the particles that, in what follows, we assume to be unity without loss of generality. We shall consider the initial value problem corresponding to

$$f(0,x,v) \;=\; f_0(x,v) \ge 0\ . \tag{1.4}$$

This system is called the *Vlasov–Poisson system for charged particles*. The main feature we add to standard versions of the Vlasov–Poisson system is an external

*Received by the editors November 16, 2001; accepted for publication (in revised form) May 8, 2002; published electronically December 3, 2002. This work was supported by the EU-TMR "Asymptotic Methods in Kinetic Theory" project ERBFRMXCT 970157 and by the Spanish DGI-MCYT project BFM 2002-01710.

http://www.siam.org/journals/sima/34-2/39843.html

†Departamento de Matemática Aplicada, Universidad de Granada, 18071 Granada, Spain (caceresg@ugr.es, carrillo@ugr.es).

‡Ceremade, Université Paris-Dauphine, Place de Lattre de Tassigny, 75775 Paris Cedex 16, France (dolbeaul@ceremade.dauphine.fr).

potential that confines particles and allows the existence of steady states. For this reason, we will refer to $\phi_e(x)$ as a *confinement potential*.

The aim of this paper is to establish the nonlinear stability of special stationary solutions in $L^p(\mathbb{R}^6)$ with $p \in [1,2]$ and explicit constants, at least in some cases (see section 3). For this purpose, we shall use an entropy, which is also called Casimir-energy, free energy, relative entropy, or Lyapunov functional in the literature. The stationary solution is a minimizer, under constraints, of the entropy; or, reciprocally, the entropy functional is determined by the shape in energy of the stationary solution. Our first main result corresponds to a p which is fixed by the entropy.

THEOREM 1.1. *Let ϕ_e be a bounded-from-below function on $\mathbb{R}^3$ with $\phi_e(x) \to \infty$ as $|x| \to +\infty$ such that $(x,s) \mapsto s^{3/2-1}\gamma(s+\phi_e(x))$ belongs to $L^1 \cap L^\infty(\mathbb{R}^3, L^1(\mathbb{R}))$. Here γ is the inverse of $-\sigma'$, eventually extended by 0, where σ is a bounded-from-below and strictly convex function of class C^2.*

Let f be a weak solution of the Vlasov–Poisson system corresponding to a nonnegative initial data f_0 in $L^1 \cap L^{p_0}$, $p_0 = (12+3\sqrt{5})/11$, such that $\sigma(f_0)$ and $(|\phi_e|+|v|^2)f_0$ belong to $L^1(\mathbb{R}^6)$. If $\inf_{s\in(0,+\infty)} \sigma''(s)/s^{p-2} > 0$ for some $p \in [1,2]$, then there exists an explicit constant $C > 0$, which depends only on f_0, such that for any $t > 0$, $f = f(t)$ satisfies

$$\|f-f_\infty\|_{L^p}^2 \le C \int_{\mathbb{R}^6} [\sigma(f_0)-\sigma(f_\infty)-\sigma'(f_\infty)(f_0-f_\infty)]\, d(x,v) + \frac{1}{2}\int_{\mathbb{R}^3} |\nabla(\phi_0-\phi_\infty)|^2\, dx\,,$$

where $\big(f_\infty(x,v) = \gamma(\frac{1}{2}|v|^2 + \phi_e(x) + \phi_\infty(x)), \phi_\infty\big)$ is a stationary solution of the Vlasov–Poisson system and ϕ_0 is given by (1.3) *at $t = 0$.*

The value of p_0 arises from the paper [34] by Hörst and Hunze in order to define weak solutions (see section 2 for more details). Note that some of our results can be extended to weaker notions of solutions, like the renormalized solutions introduced by DiPerna and Lions in [26], as we shall see later.

Also, let us point out that assumptions over σ in Theorem 1.1 can be translated into assumptions over γ if needed. We remark that our stationary states are obtained as minimizers of entropy functionals; thus hypotheses over σ are more natural.

Our second main result is a stability result in L^2, which can be written as follows in the case of Maxwellian stationary solutions.

THEOREM 1.2. *Under the same assumptions as in Theorem* 1.1*, except that we assume now $p_0 = 2$ and $\sigma(s) = s\log s - s$, there exists a convex functional $\mathcal{F}$ reaching its minimum at $f = f_\infty$ such that any weak solution to* (1.1)–(1.4) *satisfies*

$$\|f(t,\cdot) - f_\infty\|_{L^2}^2 \le \mathcal{F}[f_0]\,.$$

With the notation of Theorem 1.1, $p = 1$, $\gamma(s) = e^{-s}$, and (f_∞, ϕ_∞) is given by $f_\infty(x,v) = \frac{e^{-|v|^2/2}}{(2\pi)^{3/2}}\,\rho_\infty(x)$ with $-\Delta\phi_\infty = \rho_\infty = \|f_0\|_{L^1}\frac{e^{-(\phi_\infty+\phi_e)}}{\int e^{-(\phi_\infty+\phi_e)}\,dx}$. More general statements will be given in the rest of the paper.

Theorem 1.1 is based on a somehow canonical method to relate entropies and special stationary solutions, at least for $p = 1$ or $p = 2$. Here we get an L^p-nonlinear stability result, $1 \le p \le 2$, for a whole family of stationary solutions. It is also possible to take advantage of the uniform boundedness of the stationary solution to introduce new possible choices of the entropy functional and get stability results in L^q with $q \ne p$: for instance, $q = 2$ and $p = 1$ in Theorem 1.2. Note that Theorem 1.2 provides an L^2-stability result for the Maxwellian stationary solutions, which is not included in Theorem 1.1 (see section 4).

Similar ideas have been used previously in various contexts: for gravitational systems (without confinement) in [42, 44, 30, 31, 32] using the Casimir-energy method, and for systems in bounded domains in [6, 7], using entropy fluxes involving Darrozès–Guiraud-type estimates. For confinement, we shall refer to [27] and also to [11, 24, 10] in the case of models with a Fokker–Planck term. Entropy methods have recently been adapted to nonlinear diffusions: see, for instance, [2] in the linear case and [13, 14, 20, 21, 39, 23, 22] in the nonlinear case, with applications to models where a Poisson coupling is involved [2, 8, 9] (also see references therein for earlier works). The estimates of Csiszár–Kullback type are indeed exactly the same in kinetic and parabolic frameworks.

In the electrostatic case of the Vlasov–Poisson system, the most relevant reference for our paper is [12] (also see [4, 5, 29] for earlier results in plasma physics). In [12], Braasch, Rein, and Vukadinović consider compactly supported classical solutions to the Cauchy problem and stationary solutions which are compactly supported in the energy variable and depending on additional invariants of the particle motion. The scope of our paper is to extend their approach to general weak solutions and to emphasize the interplay of the regularity of the initial data and the various possible functionals and norms. We improve and complement results in [12] in several ways. We generalize stationary states in two directions: (1) We allow them to be not compactly supported in energy variable (Maxwellian stationary states), and (2) the dependence on energy and on other invariants of motion includes states which have not been factorized (see section 6 for details). Theorems 1.1 and 1.2 are valid for either weak or renormalized solutions (see below for details). And finally, we obtain stability bounds in L^q spaces $1 \le q \le 2$ (while in [12] only for $q = 2$).

We are going to work in the framework of weak [34, 36] or renormalized [26, 38] solutions, which of course contains the case of classical solutions. As we shall see below, there is a natural class of stationary solutions and L^p norms with respect to which the stability can be studied, but we will also consider other L^q norms. For instance, Maxwellian steady states are known to be asymptotically stable in $L^1(\mathbb{R}^6)$ for the Vlasov–Poisson–Fokker–Planck (VPFP) system [11, 10, 27, 24]. It turns out that they are stable for the Vlasov–Poisson system, in L^1 of course, but also in other norms. This question initially motivated our study and has been used to extend [12] (see Theorem 1.2).

This paper is organized as follows. We start our discussion with an overview of the definitions and properties of the solutions to the Vlasov–Poisson system. We also introduce in section 2 the family of stationary solutions we are dealing with and some of their properties. Section 3 contains the proof of a generalized version of Theorem 1.1. Theorem 1.2 is proved in section 4. In section 5, we establish some relations among various nonlinear stability results and generalize Theorem 1.2. Finally, in section 6 we consider more general steady states depending on additional invariants, for which we prove an extension of Theorem 1.2.

2. Notions of solution and stationary solutions.

2.1. Weak and renormalized solutions to the Cauchy problem. A *classical solution* [41, 43, 33, 28] is a solution to the Cauchy problem (1.1)–(1.4) for which the derivatives hold in the classical sense and the force term F satisfies a Lipschitz condition. Our approach applies to weaker notions of solutions. By *weak solution* [3, 34, 36], we mean a solution in the distributional sense, for which the force field F is not smooth enough to apply the classical characteristics theory (see below for a precise definition). Essentially, we are going to use the framework of *weak solutions* ($\mathcal{W}$)

of Hörst and Hunze [34] and, as a special case, that of Lions and Perthame [36], for which further interpolations identities are available. These last solutions are sometimes called *strong solutions* [40], and we shall denote them by $(\mathcal{S})$. For solutions corresponding to initial data with very low regularity, we shall use the *renormalized solutions* $(\mathcal{R})$ of DiPerna and Lions [26, 38].

Before making these notions of solution precise, let us introduce some notation and a basic hypothesis on the initial data. We shall refer to the Cauchy problem for the Vlasov–Poisson system with initial data f_0 as the *initial value problem* (1.1)–(1.4). We assume

(H1) $\qquad f_0$ *is a nonnegative function in* $L^1(\mathbb{R}^6)$

and denote by $M := \|f_0\|_{L^1}$ its *mass*. Let ϕ_0 be the solution to the Poisson equation at $t = 0$, corresponding to $f = f_0$ in (1.3).

Throughout this paper, we consider global in time solutions: $\mathbb{R}_0^+ = [0, \infty)$ is the time interval. As a preliminary step, we can state the following result (see the appendix for a proof).

PROPOSITION 2.1. *For any nonnegative function* f_0 *in* $L^1(\mathbb{R}^6)$*, there exists a nonnegative strictly convex function* σ *such that* $\lim_{s\to+\infty} \sigma(s)/s = +\infty$ *and* $\sigma(f_0) \in L^1(\mathbb{R}^6)$.

To obtain stability results, we are going to impose further constraints on σ, which will be strongly related to the choice of the entropy or to the choice of a special stationary solution. However, we first have to define a precise notion of solution.

DEFINITION 2.2. *Let* $p \in [1, \infty]$. *A function* $f \in L^\infty(\mathbb{R}_0^+, L^p(\mathbb{R}^6))$ *is a global* weak solution *of* (1.1)–(1.4) *with initial data* f_0 *if and only if the following hold:*

1. f *is continuous on* $\mathbb{R}_0^+$ *with values in* $L^s(\mathbb{R}^6)$*, where* $s \in [1, p)$ *(*$s = 1$ *if* $p = 1$*), with respect to the* $\sigma(L^p, L^{p'})$ *topology (weak topology for* $p < \infty$ *and weak* $*$ *topology for* $p = \infty$*). Here* p *and* p' *are the Hölder conjugates.*
2. $f(0, \cdot) = f_0$.
3. *The function* $(x, v) \mapsto f(t, x, v)F(t, x)$ *is locally integrable over* $\mathbb{R}^6$ *for all* $t \geq 0$*. (Since* $f(t) \in L^1(\mathbb{R}^6)$ *for any fixed* t*,* $F(t, \cdot)$ *is defined almost everywhere on* $\mathbb{R}^3$ *and is locally integrable.)*
4. *For all test functions* $\chi \in {C_c}^1(\mathbb{R}^6)$*, the function* $\varrho(t) := \int \chi(x, v) f(t, x, v)\, d(x, v)$ *is continuously differentiable on* $\mathbb{R}_0^+$ *and*

$$\varrho'(t) = \int v \cdot \nabla_x \chi(x, v)\, f(t, x, v)\, d(x, v) + \int F(t, x) \cdot \nabla_v \chi(x, v)\, f(t, x, v)\, d(x, v)\,.$$

Note that a weak solution for $p > 1$ is a weak solution for all $q \in [1, p]$. According to Hörst and Hunze [34], such *weak solutions* exist in the case $\phi_e \equiv 0$ globally in time if we assume that f_0 satisfies

$(\mathcal{W})$ $f_0 \geq 0$, $f_0 \in L^1(\mathbb{R}^6) \cap L^p(\mathbb{R}^6)$, $p \geq p_0 = (12 + 3\sqrt{5})/11 = 1.70075\ldots$, and

$$\int_{\mathbb{R}^6} \Big(|v|^2 + \phi_e(x)\Big) f_0(x, v)\, d(x, v) < \infty\,.$$

We shall also consider the subcase of the so-called *strong solutions* of Lions and Perthame [36]:

$(\mathcal{S})$ $f_0 \geq 0$, $f_0 \in L^1(\mathbb{R}^6) \cap L^\infty(\mathbb{R}^6)$, and for some $m > 3$,

$$\int_{\mathbb{R}^6} \Big(|v|^m + \phi_e(x)\Big) f_0(x, v)\, d(x, v) < \infty\,.$$

Remark 2.3. In case $(\mathcal{W})$, $\nabla\phi_0 \in L^2(\mathbb{R}^3)^3$ [34] as a consequence of the interpolation inequality, $\|\rho\|_{L^q} \le C\,\|f\|_{L^p}^{\theta}\,\|\,|v|^2 f\|_{L^1}^{1-\theta}$ with $q = \frac{5p-3}{3p-1}$, $\theta \in (0,1)$; and of the Hardy–Littlewood–Sobolev inequality, $\|\nabla\phi\|_{L^r} \le C\,\|\rho\|_{L^q}$ with $\frac{1}{q} - \frac{1}{r} = \frac{1}{3}$. The case $p = p_0$ is obtained by imposing $r = p'$.

Without assumptions on the initial energy, it is still possible to give global existence results [15, 16]. Also note that if $(\mathcal{W})$ is satisfied, $f_0 \log f_0 \in L^1(\mathbb{R}^6)$, as we shall see in section 4, provided $e^{-\beta\,\phi_e} \in L^1(\mathbb{R}^3)$ for some $\beta > 0$.

In this paper, we will also consider weaker notions of solutions.

DEFINITION 2.4. *Assume that*

$(\mathcal{R})$ *f_0 is a nonnegative function in $L^1(\mathbb{R}^6)$ such that $f_0 \log f_0 \in L^1(\mathbb{R}^6)$ and*

$$\int_{\mathbb{R}^6} \Big(\frac{1}{2}|v|^2 + \phi_e(x)\Big) f_0(x,v)\; d(x,v) + \frac{1}{2}\int_{\mathbb{R}^3} |\nabla_x \phi_0|^2\; dx < \infty\;.$$

We shall say that $f \in C^0(\mathbb{R}_0^+, L^1(\mathbb{R}^6))$ is a renormalized solution *of* (1.1)–(1.4) *on $\mathbb{R}_0^+$ with initial data f_0 if and only if*

1. *the quantities*

$$\int_{\mathbb{R}^6} \Big(\frac{1}{2}|v|^2 + \phi_e(x) + \phi(x,t)\Big)\, f(t,x,v)\; d(x,v)$$
$$\text{and}\ \int_{\mathbb{R}^6} f(t,x,v)\,\log f(t,x,v)\; d(x,v)$$

   are bounded from above, uniformly in $t \ge 0$;

2. *$\beta(f) = \log(1+f)$ is a weak solution of*

$$\frac{\partial}{\partial t}\beta(f) + v\cdot\nabla_x\beta(f) + F(t,x)\cdot\nabla_v\beta(f) = 0$$

   considered in the distributional sense, where F is defined according to (1.2) *and* (1.3).

In the case in which $e^{-\beta\,\phi_e} \in L^1(\mathbb{R}^3)$ for some $\beta > 0$, weak solutions for $p > 1$ are also renormalized solutions (see Lemma 4.1).

PROPOSITION 2.5. *Let f_0 verify $(\mathcal{R})$ and assume that ϕ_e is a nonnegative potential such that $\lim_{|x|\to+\infty} \phi_e(x) = +\infty$. If ϕ_e is in $W^{1,1}_{loc}(\mathbb{R}^3)$, then* (1.1)–(1.4) *admits a global in time renormalized solution. If, moreover, ϕ_e belongs to $W^{1,q}_{loc}$ for $q \ge \frac{5p-3}{2(p-1)}$ and if $(\mathcal{W})$ holds, then* (1.1)–(1.4) *admits a weak solution.*

Proof. This result can be obtained by adapting the proofs of [34, 36, 26, 38]. For renormalized solutions, characteristics can be defined according to [25, 35] as soon as ϕ_e is in $W^{1,1}_{loc}(\mathbb{R}^3)$. Details are left to the reader. □

Weak or renormalized solutions have the following properties:

1. The distribution function is nonnegative for all $t \ge 0$.
2. Conservation of mass: for any $t \ge 0$,

$$\int_{\mathbb{R}^6} f(t,x,v)\; d(x,v) \;=\; \int_{\mathbb{R}^6} f_0(x,v)\; d(x,v) \;=\; M\;.$$

3. Finite kinetic energy, potential energy, and entropy: for any $t \ge 0$,

$$\int_{\mathbb{R}^6} \left(\frac{1}{2}|v|^2 + \phi_e(x) + \phi(x)\right) f\; d(x,v) \le \int_{\mathbb{R}^6} \left(\frac{1}{2}|v|^2 + \phi_e(x) + \phi_0(x)\right) f_0\; d(x,v)$$
$$\text{and}\quad \int_{\mathbb{R}^6} f\,\log f\; d(x,v) \le \int_{\mathbb{R}^6} f_0\,\log f_0\; d(x,v)\;,$$

with equality in the case of classical solutions (see Corollary 2.8 for an application).

4. In case $(\mathcal{S})$, for any $t \geq 0$,

$$\|f(t,\cdot)\|_{L^\infty(\mathbb{R}^6)} \leq \|f_0\|_{L^\infty(\mathbb{R}^6)} .$$

5. Moreover, if we assume that

$$\int_{\mathbb{R}^6} \sigma(f_0)\ d(x,v) < \infty \tag{H2}$$

for some strictly convex continuous function $\sigma : \mathbb{R}_0^+ \to \mathbb{R}$, then for any $t \geq 0$,

$$\int_{\mathbb{R}^6} \sigma(f)\ d(x,v)\ \leq\ \int_{\mathbb{R}^6} \sigma(f_0)\ d(x,v)\ ,$$

with equality in the case of classical solutions (see Corollary 2.8 for an application).

2.2. Stationary solutions and entropy functionals. Let us introduce further notation. For any function $f \in L^1(\mathbb{R}^6)$, let $\phi = \phi[f]$ be the solution of $-\Delta\phi = \int_{\mathbb{R}^3} f\ dv$ in $L^{3,\infty}(\mathbb{R}^3)$ given by the convolution with the Green function of the Laplacian. The operator ϕ is linear and satisfies

$$\int_{\mathbb{R}^6} f\,\phi[g]\ d(x,v) = \int_{\mathbb{R}^6} g\,\phi[f]\ d(x,v)\ .$$

Any function $f_{\infty,\sigma}$ such that

$$f_{\infty,\sigma}(x,v) = \gamma\left(\frac{1}{2}|v|^2 + \phi[f_{\infty,\sigma}](x) + \phi_e(x) - \alpha\right) \tag{2.1}$$

is a stationary solution of the Vlasov–Poisson system. Such a solution exists if and only if

$$-\Delta\phi_{\infty,\sigma} = G_\sigma(\phi_{\infty,\sigma} + \phi_e - \alpha) \quad \text{with} \quad G_\sigma(\phi) = 4\pi\sqrt{2}\int_0^{+\infty} \sqrt{s}\ \gamma(s+\phi)\ ds$$

has a solution $\phi_{\infty,\sigma} = \phi[f_{\infty,\sigma}]$ such that $\int_{\mathbb{R}^6} f_{\infty,\sigma}\ d(x,v) = M$. The constant α is therefore determined by the total mass M. Under assumptions that we are going to specify now, we will prove that such a stationary solution exists and is unique (see Lemma 2.7).

Let us consider σ such that γ is the generalized inverse of $-\sigma'$ (eventually extended by 0): σ is convex (resp., strictly convex) if and only if γ is monotone nonincreasing (resp., decreasing in its support). With these notations, we assume that σ and ϕ_e verify the following:

(H3) $\sigma \in C^2(\mathbb{R}^+) \cap C^0(\mathbb{R}_0^+)$ *is a bounded-from-below strictly convex function such that*

$$\lim_{s\to+\infty} \frac{\sigma(s)}{s} = +\infty\ ;$$

(H4) $\phi_e : \mathbb{R}^3 \to \mathbb{R}$ *is a measurable bounded-from-below function such that*

$$\lim_{|x|\to+\infty} \phi_e(x) = +\infty$$

and $x \mapsto G_\sigma(\phi_e(x)) = 4\pi\sqrt{2}\int_0^{+\infty} \sqrt{s}\ \gamma(s+\phi_e(x))\ ds$ *belongs to* $L^1 \cap L^\infty(\mathbb{R}^3)$.

The conditions on the growth of ϕ_e and on the decay of γ will be referred to as *confinement conditions*. We are going to adapt the proofs given in [27] for the case $\gamma(s) = e^{-s}$ and in [6, 7] for the bounded domain case to prove the existence of a stationary solution $f_{\infty,\sigma}$. The existence of $\alpha = \alpha(M)$ will be a consequence of the proof.

Let $M > 0$ and consider on $L^1_M(\mathbb{R}^6) = \{f \in L^1(\mathbb{R}^6) \ : \ f \geq 0 \text{ a.e., } \|f\|_{L^1} = M\}$ the functional

$$K_\sigma[f] = \int_{\mathbb{R}^6} \left[\sigma(f) + \left(\frac{1}{2}|v|^2 + \phi_e(x) \right) f \right] \, d(x,v) + \frac{1}{2} \int_{\mathbb{R}^3} |\, \nabla\phi[f]\,|^2 \, dx \, .$$

DEFINITION 2.6. *Given f and g in $L^1_M(\mathbb{R}^6)$, the relative entropy of f with respect to g is*

$$\Sigma_\sigma[f|g] := \int_{\mathbb{R}^6} [\sigma(f) - \sigma(g) - \sigma'(g)(f-g)]\, d(x,v) + \frac{1}{2} \int_{\mathbb{R}^3} |\nabla\phi[f-g]|^2 \, dx \tag{2.2}$$

LEMMA 2.7. *Under assumptions* (H3)–(H4)*, K_σ is a strictly convex bounded-from-below functional on $L^1_M(\mathbb{R}^6)$. It has a unique global minimum, $f_{\infty,\sigma}$, which takes the form* (2.1) *and is therefore a stationary solution of the Vlasov–Poisson system. Moreover $\Sigma_\sigma[f|f_{\infty,\sigma}]$ can be written as*

$$\Sigma_\sigma[f|f_{\infty,\sigma}] = K_\sigma[f] - K_\sigma[f_{\infty,\sigma}] \, . \tag{2.3}$$

and $\sigma(f_{\infty,\sigma})$ and $\sigma'(f_{\infty,\sigma})f_{\infty,\sigma}$ belong to $L^1(\mathbb{R}^6)$.

Proof. Assumption (H4) gives that $K_\sigma[f]$ is bounded from below by Jensen's inequality. By hypothesis (H3) K_σ is convex, so we may pass to the limit in a minimizing sequence involving the semicontinuity property. The limit $f_{\infty,\sigma}$ belongs to $L^1_M(\mathbb{R}^6)$ because of the Dunford–Pettis criterion. Equation (2.1) is the corresponding Euler–Lagrange (where α enters as the Lagrange multiplier associated to the constraint on the L^1 norm). Identity (2.2) easily follows by a direct computation using (2.1). ∎

Note that $\Sigma_\sigma[f|f_{\infty,\sigma}]$ is obviously nonnegative, since $K_\sigma[f]$ attains its unique minimum at $f = f_{\infty,\sigma}$.

COROLLARY 2.8. *Consider a renormalized or weak solution f of* (1.1)–(1.4) *under assumptions* (H1), (H2), (H3), *and* (H4). *Then $\Sigma_\sigma[f(t)|f_{\infty,\sigma}] \leq \Sigma_\sigma[f_0|f_{\infty,\sigma}]$.*

The proof relies on standard semicontinuity arguments and is left to the reader.

Example 2.9. (1) Let $\sigma_q(s) = s^q$, with $\gamma_q(s) = (-s/q)_+^{1/(q-1)}$, for some given $q > 1$. With the notation $f_{\infty,q} = f_{\infty,\sigma_q}$ and $\phi_{\infty,q} = \phi[f_{\infty,\sigma_q}]$, this stationary solution satisfies the nonlinear Poisson equation

$$-\Delta\phi_{\infty,q} = C_q \, (\alpha(M) - \phi_e - \phi_{\infty,q})_+^{\frac{3}{2} + \frac{1}{q-1}} \ ,$$

where $C_q = (2\pi)^{3/2} \, q^{-\frac{1}{q-1}} \, \Gamma(\frac{q}{q-1})/\Gamma(\frac{5q-3}{2(q-1)})$.

(2) The limit case as $q \to 1$ corresponds to $\sigma_1(s) = s\log s - s$ and $\gamma_1(s) = e^{-s}$. In this case we obtain the Maxwellian stationary solution

$$f_{\infty,1}(x,v) \ = \ m(x,v) \ = \ M \, \frac{e^{-\frac{1}{2}|v|^2}}{(2\pi)^{3/2}} \, \frac{e^{-(\phi_{\infty,1}(x)+\phi_e(x))}}{\int_{\mathbb{R}^3} e^{-(\phi_{\infty,1}(x)+\phi_e(x))} \, dx} \, , \tag{2.4}$$

where $\phi_{\infty,1}$ is given by the Poisson–Boltzmann equation

$$-\Delta_x \phi_{\infty,1} \ = \ \int_{\mathbb{R}^3} m(x,v) \, dv \ = \ M \, \frac{e^{-(\phi_{\infty,1}+\phi_e)}}{\int_{\mathbb{R}^3} e^{-(\phi_{\infty,1}+\phi_e)} \, dx} \, . \tag{2.5}$$

(3) A less standard case is given by

$$\sigma(t) = \begin{cases} 2\int_1^{\sqrt{-\log t}} s^2 e^{-s^2}\ ds & \text{if } 0 < t \le 1\ , \\ 0 & \text{if } t > 1\ , \end{cases}$$

which corresponds to $\gamma(t) = e^{-t^2}$.

In the following sections, the various cases of this example will be analyzed. They will motivate a more general treatment. For simplicity, we shall write $\Sigma_q[f|f_{\infty,q}]$ instead of $\Sigma_{\sigma_q}[f|f_{\infty,\sigma_q}]$ for $q \ge 1$.

3. L^p-nonlinear stability. In this section, we give an L^p-nonlinear stability result for $f_{\infty,\sigma}$, $1 \le p \le 2$, with minimal convexity assumptions on the initial data and an explicit stability constant. It is based on the following result.

PROPOSITION 3.1. *Let f and g be two nonnegative functions in $L^1(\mathbb{R}^6) \cap L^p(\mathbb{R}^6)$, $p \in [1,2]$, and consider a strictly convex function $\sigma : \mathbb{R}_0^+ \to \mathbb{R}$ in $C^2(\mathbb{R}^+) \cap C^0(\mathbb{R}_0^+)$. Let $A = \inf\left\{\sigma''(s)/s^{p-2} :\ s \in (0,\infty)\right\}$. If $A > 0$, then the following inequality holds:*

$$\Sigma_\sigma[f|g] \ge 2^{-2/p} A \left[\max\left(\|f\|_{L^p}^{2-p}, \|g\|_{L^p}^{2-p}\right)\right]^{-1} \|f-g\|_{L^p}^2 + \frac{1}{2}\int_{\mathbb{R}^3} |\nabla_x(\phi[f]-\phi[g])|^2\ dx\ . \tag{3.1}$$

Proof. The case $p = 1$ is the well-known Csiszár–Kullback inequality (see, for instance, [1]) that we are going to adapt to the case $p \ge 1$.

Assume first that $f > 0$. By a Taylor development at order 2 of σ we deduce that we can write the relative entropy for f and g as

$$\begin{aligned} \Sigma_\sigma[f|g] &= \frac{1}{2}\int_{\mathbb{R}^6} \sigma''(\xi)|f-g|^2\ d(x,v) + \frac{1}{2}\int_{\mathbb{R}^3} |\nabla_x(\phi[f]-\phi[g])|^2\ dx \\ &\ge \frac{A}{2}\int_{\mathbb{R}^6} \xi^{p-2}|f-g|^2\ d(x,v) + \frac{1}{2}\int_{\mathbb{R}^3} |\nabla_x(\phi[f]-\phi[g])|^2\ dx\ , \end{aligned} \tag{3.2}$$

where ξ lies between f and g. If $p = 2$, the result is obvious. Let $1 \le p < 2$. By Hölder's inequality, for any $h > 0$ and for any measurable set $\mathcal{A} \subset \mathbb{R}^6$, we get

$$\int_{\mathcal{A}} |f-g|^p\, h^{-\alpha}\, h^\alpha\ d(x,v) \le \left(\int_{\mathcal{A}} |f-g|^2\, h^{p-2}\ d(x,v)\right)^{p/2} \left(\int_{\mathcal{A}} h^{\alpha s}\ d(x,v)\right)^{1/s}$$

with $\alpha = p\,(2-p)/2$, $s = 2/(2-p)$. Thus

$$\left(\int_{\mathcal{A}} |f-g|^2\, h^{p-2}\ d(x,v)\right)^{p/2} \ge \left(\int_{\mathcal{A}} |f-g|^p\ d(x,v)\right)\ \left(\int_{\mathcal{A}} h^p\ d(x,v)\right)^{(p-2)/2}\ .$$

We apply this formula to two different sets.

(i) On $\mathcal{A} = \mathcal{A}_1 = \{(x,v) \in \mathbb{R}^6 : f(x,v) > g(x,v)\}$, use $\xi^{p-2} > f^{p-2}$ and take $h = f$:

$$\left(\int_{\mathcal{A}_1} |f-g|^2 \xi^{p-2}\ d(x,v)\right)^{p/2} \ge \left(\int_{\mathcal{A}_1} |f-g|^p\ d(x,v)\right)\ \|f\|_{L^p}^{-(2-p)\,p/2}\ .$$

(ii) On $\mathcal{A} = \mathcal{A}_2 = \{(x,v) \in \mathbb{R}^6 : f(x,v) \le g(x,v)\}$, use $\xi^{p-2} \ge g^{p-2}$ and take $h = g$:

$$\left(\int_{\mathcal{A}_2} |f-g|^2 \xi^{p-2}\ d(x,v)\right)^{p/2} \ge \left(\int_{\mathcal{A}_2} |f-g|^p\ d(x,v)\right)\ \|g\|_{L^p}^{-(2-p)\,p/2}\ .$$

To prove (3.1) in the case $f > 0$, we just add the two previous inequalities in (3.2) and use the inequality $(a+b)^r \le 2^{r-1}(a^r+b^r)$ for any $a, b \ge 0$ and $r \ge 1$. To handle the case $f \ge 0$, we proceed by a density argument: apply (3.1) to $f_\epsilon(x,v) = f(x,v) + \epsilon\, e^{-|x|^2-|v|^2}$ and let $\epsilon \to 0$ using Lebesgue's convergence theorem. □

This proposition can be applied to weak or renormalized solutions, thus proving the first main result of this paper, which is a more detailed version of Theorem 1.1.

THEOREM 3.2. *Let f_0 verify* (H1), (H2), *and either* $(\mathcal{R})$ *or* $(\mathcal{W})$. *Assume* (H3) *and* (H4). *If f is a weak or renormalized solution of* (1.1)–(1.4) *with initial value f_0, then*

$$\|\nabla\phi - \nabla\phi_{\infty,\sigma}\|^2_{L^2} \quad \le \quad 2\,\Sigma_\sigma[f_0|f_{\infty,\sigma}]\,.$$

Assume that $A = \inf\left\{\sigma''(s)/s^{p-2} \ :\ s \in (0,\infty)\right\}$ *is positive for some* $p \in [1,2]$. *If* $p = 1$, *assume moreover that* $e^{-\phi_e} \in L^1$. *Then* $f_0 \in L^p(\mathbb{R}^6)$ *and*

$$\|f(t) - f_{\infty,\sigma}\|^2_{L^p} \le C(f_0,\sigma)\,\Sigma_\sigma[f_0|f_{\infty,\sigma}]$$

for any $t \ge 0$, *where* $C(f_0,\sigma)$ *is a constant, which takes the explicit form*

$$C(f_0,\sigma) = \begin{cases} \frac{2^{2/p}}{A}\max\left(\|f_0\|^{2-p}_{L^p}, \|f_{\infty,\sigma}\|^{2-p}_{L^p}\right) & \text{if } p > 1\,, \\ \\ \frac{4}{A}M & \text{if } p = 1\,. \end{cases}$$

In case $(\mathcal{S})$, $C(f_0,\sigma)$ *is also bounded by* $\frac{2^{2/p}}{A}\,M^{(2-p)/p}\,\mathcal{M}^{(2-p)(p-1)/p}$ *with* $\mathcal{M} = \max\left(\|f_0\|_{L^\infty}, \|f_{\infty,\sigma}\|_{L^\infty}\right)$.

Proof. The proof is a straightforward consequence of Lemma 2.7, Corollary 2.8, and Proposition 3.1 once it is known that $C(f_0,\sigma)$ is finite. Although we directly prove an estimate of $\|f(t) - f_{\infty,\sigma}\|^2_{L^p}$ in terms of $\Sigma_\sigma[f_0|f_{\infty,\sigma}]$, we may notice that, for $p > 1$, two integrations give the inequality

$$\sigma(s) - \sigma(s_0) - \sigma'(s_0)\,(s - s_0) \ge \frac{A}{p\,(p-1)}\left[s^p - s_0^p - p\,s_0^{p-1}\,(s - s_0)\right]$$

for any $(s, s_0) \in (0+\infty)^2$. Applied to f and $f_{\infty,\sigma}$, this means that on $\mathbb{R}^6$

$$\sigma(f) - \sigma(f_{\infty,\sigma}) - \sigma'(f_{\infty,\sigma})\,(f - f_{\infty,\sigma}) \ge \frac{A}{p\,(p-1)}\left[f^p - f^p_{\infty,\sigma} - p\,f^{p-1}_{\infty,\sigma}\,(f - f_{\infty,\sigma})\right]\,, \tag{3.3}$$

which proves that f belongs to $L^\infty(\mathbb{R}^+, L^p(\mathbb{R}^6))$ (by $\|f_0\|_{L^p}$, according to Corollary 2.8 applied with $\sigma(s) = \sigma_p(s) = s^p$). The constant $C(f_0,\sigma)$ involves $\|f_0\|_{L^p}$, which is therefore itself bounded in terms of $\sigma(f_0)$ and $f_0\,\sigma'(f_0)$.

If $p = 1$, the condition that $e^{-\phi_e} \in L^1$ shows that $f_{\infty,\sigma}$ also belongs to L^1. In that case, inequality (3.3) is replaced by

$$\sigma(f) - \sigma(f_{\infty,\sigma}) - \sigma'(f_{\infty,\sigma})\,(f - f_{\infty,\sigma}) \ge A\left[f\log\left(\frac{f}{f_{\infty,\sigma}}\right) - (f - f_{\infty,\sigma})\right]\,.$$

The details of the proof are left to the reader. □

Remark 3.3. Note that $A = p\,(p-1)$ if $\sigma = \sigma_p$, $p > 1$, and $A = 1$ if $p = 1$ and $C(f_0,\sigma_2) = 1$. The expression of $C(f_0,\sigma)$ is optimal at least for $\sigma = \sigma_p$ in the limit $\|f_0 - f_{\infty,\sigma}\|_{L^p} \to 0$ (see [1] for a discussion in the case $p = 1$).

For $p > 2$, Hölder's inequality holds in the reverse sense: $\|f(t) - f_{\infty,\sigma}\|^2_{L^p} + \|\nabla\phi - \nabla\phi_{\infty,\sigma}\|^2_{L^2}$ controls $\Sigma_\sigma[f_0|\sigma]$.

For $p = 1$, we recover the classical Csiszár–Kullback inequality in Proposition 3.1 and a stability result in L^1 (see [1, 2]) which is natural in the framework of renormalized solutions (if $f\log f$ belongs to L^1: see Lemma 4.1 below).

4. L^2-nonlinear stability of Maxwellian steady states. In [12], Braasch, Rein, and Vukadinović introduce modified Lyapunov functionals for proving L^2-stability for certain steady states (see section 5 for more details). In this section, we shall extend this approach to the Maxwellian case. The main idea is the following: Although $\sigma''(s) = 1/s$ is not bounded from below uniformly away from 0 (which would be the condition to apply directly Proposition 3.1 in L^2), since $f_{\infty,1}$ is bounded in L^∞ by a constant $\bar{s}$, it is sufficient to consider the infimum of σ'' in $(0, \bar{s})$.

In the Maxwellian case, we first notice that (H2) follows from the other assumptions.

LEMMA 4.1. *Assume that $e^{-\beta\,\phi_e}$ belongs to $L^1(\mathbb{R}^3)$ for some $\beta > 0$. Let f be a nonnegative function in $L^1 \cap L^q(\mathbb{R}^6)$, $q > 1$, such that $(x,v) \mapsto (|v|^2 + \phi_e(x))f(x,v) \in L^1(\mathbb{R}^6)$. Then $f \log f$ belongs to $L^1(\mathbb{R}^6)$.*

Proof. Depending on the sign of $\log f$, we are going to consider two cases.

(1) Define $g(x,v) = e^{-\frac{\beta}{2}|v|^2 - \beta\,\phi_e(x)}$. On $\mathcal{A} = \{(x,v) \in \mathbb{R}^6 \ : \ f(x,v) < 1\}$, using Jensen's inequality, we get

$$0 \geq \int_{\mathcal{A}} \left[f \log f + \beta \left(\frac{1}{2}|v|^2 + \phi_e \right) f \right] \, d(x,v) = \int_{\mathcal{A}} f \, \log\left(\frac{f}{g}\right) \, d(x,v)$$
$$\geq \|f\|_{L^1(\mathcal{A})} \log\left(\frac{\|f\|_{L^1(\mathcal{A})}}{\|g\|_{L^1(\mathcal{A})}} \right) .$$

(2) On $\mathbb{R}^6 \setminus \mathcal{A}$, we conclude using the next lemma. ∎

LEMMA 4.2. *Let f be a nonnegative function in $L^1 \cap L^q(\Omega)$, $q > 1$, for some arbitrary domain $\Omega \subset \mathbb{R}^d$, $d \geq 1$. Then*

$$\int_\Omega f(z) \, \log f(z) \, dz \leq \frac{1}{q-1} \, \|f\|_{L^1(\Omega)} \, \log\left(\frac{\|f\|^q_{L^q(\Omega)}}{\|f\|_{L^1(\Omega)}} \right) .$$

Proof. According to Hölder's inequality,

$$\|f\|^r_{L^r} \leq \|f\|_{L^1}^{\frac{q-r}{q-1}} \, \|f\|_{L^q}^{\frac{q(r-1)}{q-1}}$$

for $1 \leq r \leq q$. At $r = 1$, this is an equality and thus we may derive the inequality with respect to r at $r = 1$. ∎

Let ϕ_e and f_0 verify, respectively, (H4) for $\sigma_1(s) = s \log s - s$, and (H1), $(\mathcal{W})$. Consider a weak or renormalized solution f of (1.1)–(1.4) with initial value f_0 and the corresponding stationary solution $f_{\infty,1} = m$ given by (2.4)–(2.5). According to Theorem 3.2, m is L^1-stable:

$$\Sigma_1[f|m] \geq \frac{1}{4M} \|f - m\|^2_{L^1} \, .$$

We shall now prove an L^2-stability result for m using an appropriate cut-off functional as in [12]. Let $E_1(x,v) := \frac{1}{2}|v|^2 + \phi_{\infty,1}(x) + \phi_e(x)$. According to (H4),

$$E_{min} := \inf\{E_1(x,v) \ : \ (x,v) \in \mathbb{R}^6\} \geq \inf\{\phi_e(x) \ : \ x \in \mathbb{R}^3\} > -\infty \, .$$

Denote $m = \varphi \circ E_1$ with $\varphi(s) = \kappa \, e^{-s}$, where

$$\kappa = \frac{M}{(2\pi)^{3/2}} \left[\int e^{-\phi_{1,\infty} - \phi_e} \, dx \right]^{-1} . \tag{4.1}$$

Consider $\bar s = \varphi(E_{min})$ and define

$$\tau_1(s) := \begin{cases} s\log s - s & \text{if } s \in [0,\bar s]\,, \\ \frac{1}{2\kappa}\, e^{E_{min}}\,(s-\bar s)^2 \;-\; (E_{min} - \log\kappa)(s-\bar s) + \bar s\log\bar s - \bar s & \text{if } s \in (\bar s, +\infty)\,. \end{cases}$$

The function τ_1 is of class $C([0,\infty))\cap C^2((0,\infty))$, with $\min(\tau_1'') = e^{E_{min}}/\kappa > 0$. Since $0 \le m(x,v) \le \varphi(E_{min}) = \bar s$ for any $(x,v)\in\mathbb{R}^6$ and φ is decreasing, m is a minimizer of the modified free energy (or Casimir) functional $\Sigma_{\tau_1}[f|m] = K_{\tau_1}[f] - K_{\tau_1}[m]$, where

$$K_{\tau_1}[f] = \int_{\mathbb{R}^6} \left(\frac12\,|v|^2 + \frac12\,\phi + \phi_e\right) f\; d(x,v) \;+\; \int_{\mathbb{R}^6} \tau_1(f)\; d(x,v)\,,$$

and we can apply Theorem 3.2 with $p = 2$. This proves a refined version of Theorem 1.2. Since f belongs to L^2, $\tau_1(f)$ makes sense in L^1 according to Lemma 4.1. Let us remark that the construction of τ_1 is done in such a way that $K_{\tau_1}[m] = K_{\sigma_1}[m]$, and then Corollary 2.8 can be applied. In this framework, it is natural to work with weak rather than renormalized solutions.

THEOREM 4.3. *Assume* (H1), (H3), (H4) *for* $\sigma = \sigma_1$ *and* $(\mathcal{W})$ *for* $p = 2$. *Consider the stationary solution given by* (2.4)–(2.5). *With the above notation, every weak solution* f *of* (1.1)–(1.4) *with initial data* $f_0 \in L^1\cap L^2(\mathbb{R}^6)$ *verifies*

$$\Sigma_{\tau_1}[f_0|m] \ge \Sigma_{\tau_1}[f(t)|m] \ge \frac{1}{2\,\bar s}\,\|f(t) - m\|_{L^2}^2 \quad \textit{for all}\;\; t \ge 0\,.$$

Remark 4.4. (1) A simpler version of Theorem 4.3 holds for solutions satisfying $(\mathcal{S})$. In this case, it is not necessary to modify σ, since $\sigma_1''(s) = \frac{1}{s}$ is bounded from below in $(0, \max(\|f_0\|_{L^\infty}, \|m\|_{L^\infty})]$ by $\max(\|f_0\|_{L^\infty}, \|m\|_{L^\infty})^{-1}$.

(2) Theorem 4.3 can be generalized to any stationary solution $f_{\infty,\sigma}$ and any L^q norm with $p \ne q \in (1,2]$; see the next section.

(3) Note that in the Maxwellian case the value of κ defined by (4.1) is $e^{-\alpha(M)}$, where $\alpha = \alpha(M)$ is the constant in (2.1) which is fixed by the mass constraint.

5. General nonlinear stability results. In this section, we generalize to L^q, $1 \le q \le 2$, and to arbitrary steady states $f_{\infty,\sigma}$ the stability results of sections 3–4. We are also going to generalize the techniques used in the L^2-stability result of Braasch, Rein, and Vukadinović in [12], which can be summarized as follows. *Let* γ *be a* C^1 *function on* $\mathbb{R}$ *such that* $\gamma' < 0$ *on* $(-\infty, E_{\max})$ *and* $\gamma \equiv 0$ *on* $[E_{max}, +\infty)$ *and define* σ *as a primitive of* $-(\gamma^{-1})$, *which is well defined at least on some subinterval in* $\mathbb{R}^+$ (see, for instance, [14] for more details). *Then* $f_{\infty,\sigma}$ *is a compactly supported steady state which is* L^2*-stable among weak or renormalized solutions of* (1.1)–(1.4).

For $q > p$, the main idea is again to bound $\sigma''(s)/s^{q-2}$ from below only on the interval $(0, \bar s = \|f_{\infty,\sigma}\|_{L^\infty})$ and to modify σ on $(\bar s, +\infty)$. In this case, let us establish a useful consequence of Proposition 3.1. Let $E_\sigma(x,v) := \frac12|v|^2 + \phi_{\infty,\sigma}(x) + \phi_e(x)$ and $E_{min} := \inf\{E_\sigma(x,v) \;:\; (x,v)\in\mathbb{R}^6\}$, which is finite by assumption (H4). With the notation of sections 2–3, $f_{\infty,\sigma} = \gamma\circ(E_\sigma - \alpha)$, where α is such that $\|f_{\infty,\sigma}\|_{L^1} = M$. Take $\bar s = \gamma(E_{min} - \alpha)$ and define

$$\tau_\sigma(s) := \begin{cases} \sigma(s) & \text{if } s \in [0,\bar s]\,, \\ \psi(s) & \text{if } s \in (\bar s, +\infty) \end{cases}$$

with $\psi(s) = \frac{\sigma''(\bar s)}{\sigma_q''(\bar s)}\sigma_q(s) + (\sigma'(\bar s) - \frac{\sigma''(\bar s)}{\sigma_q''(\bar s)}\,\sigma_q'(\bar s))(s-\bar s) + \sigma(\bar s) - \frac{\sigma''(\bar s)}{\sigma_q''(\bar s)}\,\sigma_q(\bar s)$ and $\sigma_q(t) = t^q$. With the truncated Lyapunov functional $\Sigma_{\tau_\sigma}[f|f_{\infty,\sigma}] = K_{\tau_\sigma}[f] - K_{\tau_\sigma}[f_{\infty,\sigma}]$, we immediately get the following variant of Proposition 3.1.

COROLLARY 5.1. *Let f and g be two nonnegative functions in $L^1(\mathbb{R}^6) \cap L^q(\mathbb{R}^6)$, $q \in [1,2]$, and consider a strictly convex function $\sigma : \mathbb{R}_0^+ \to \mathbb{R}$ in $C^2(\mathbb{R}^+) \cap C^0(\mathbb{R}_0^+)$. With the above notation, let $B = \inf\left\{\sigma''(s)/s^{q-2} \ : \ s \in (0,\bar{s})\right\}$. If $B > 0$, then there exists a constant $C > 0$ such that*

$$\Sigma_{\tau_\sigma}[f|g] \geq C \, \|f-g\|_{L^q}^2 + \frac{1}{2}\|\nabla\phi - \nabla\phi_{\infty,\sigma}\|_{L^2}^2 \ .$$

As in the case of section 4, this estimate can be applied to get nonlinear stability results.

THEOREM 5.2. *Let f_0 verify* (H1), (H2), *and either* $(\mathcal{R})$ *or* $(\mathcal{W})$. *Assume that σ and ϕ_e satisfy* (H3) *and* (H4). *Assume that* $\inf\left\{\sigma''(s)/s^{p-2} : s \in (0,\bar{s})\right\}$ *is positive for some $p \in [1,2]$, where $\bar{s}$ is defined as above. Then $f_{\infty,\sigma}$ is L^q-nonlinearly stable among weak or renormalized solutions of* (1.1)–(1.4) *for any $q \in (1,2]$, provided $f_0 \in L^q(\mathbb{R}^6)$ if $q > p$.*

Proof. The case $q = p$ is covered by Theorem 3.2. In the case $q > p$, the proof is an easy application of Corollary 5.1: $f_{\infty,\sigma}$ is L^q-stable in the sense that there exists a constant $C > 0$ such that for any $t \geq 0$,

$$\|f(t) - f_{\infty,\sigma}\|_{L^q} \leq C \, \Sigma_{\tau_\sigma}[f_0|f_{\infty,\sigma}] \ .$$

The case $1 < q < p$ relies on Hölder's inequality and Theorem 3.2:

$$\|f(t) - f_{\infty,\sigma}\|_{L^q} \leq (2M)^{\frac{p-q}{q\,(p-1)}} \left(C(f_0,\sigma)\,\Sigma_\sigma[f_0|f_{\infty,\sigma}]\right)^{\frac{p\,(q-1)}{2q\,(p-1)}} \ . \qquad \square$$

The case $p = q = 1$ is covered by Theorem 3.2. Only the case $1 = q < p$ is left open. In the case $q > p$, notice that the L^q norm is bounded in terms of $\Sigma_{\tau_\sigma}[f_0|f_{\infty,\sigma}]$ and not in terms of $\Sigma_\sigma[f_0|f_{\infty,\sigma}]$ (as is also the case in Theorem 4.3, with $p = 1$, $q = 2$).

6. Steady states depending on additional invariants. In the previous sections, we dealt with stationary solutions depending only on the energy. Our stability analysis can be extended to steady states which depend on additional invariants of the particle motion. To avoid lengthy statements, we shall state only the generalization of Theorem 4.3. In order to emphasize the connection with the previous results, we shall abuse the same notations.

Consider the ODE system

$$\dot{X} = V \ , \quad \dot{V} = -\nabla_x\phi(t,X) - \nabla_x\phi_e(X) \ ,$$

which describes the characteristics of the Vlasov equation (1.1). We shall assume that either both ϕ and ϕ_e are locally Lipschitz (classical solutions), or both ϕ and ϕ_e are at least in $W^{1,1}_{loc}$ (using the generalized characteristics of DiPerna and Lions; see [25, 35]). A function $I : \mathbb{R}^6 \to \mathbb{R}^m$ is an *invariant of the motion* if and only if

$$\frac{d}{dt} I(X(t),V(t)) = 0$$

in an appropriate sense. Classical examples of invariants are, for instance, the angular momentum $I(x,v) = x \times v$ in the case of a central force motion (i.e., if $\phi + \phi_e$ is radially symmetric), its modulus, or one of its components: $I(x,v) \cdot \nu$, in the axisymmetric case with axis of direction $\nu \in S^2$, corresponding to a system invariant under rotations of axis ν. References on existence results of classical solutions with symmetries can be found in [28] (for stationary solutions, see [18]).

Consider stationary solutions in the form

$$f_{\infty,\sigma}(x,v) = \mu\big(E(x,v) - \alpha_M[\phi_{\infty,\sigma},\phi_e,I], I(x,v)\big)\ , \tag{6.1}$$

where α_M is a constant to be determined by $\|f_{\infty,\sigma}\|_{L^1} = M$, E is the energy, and I is an invariant of the motion. Note that E depends on $\phi_{\infty,\sigma} = \phi[f_{\infty,\sigma}]$. For simplicity, we suppose that I is a scalar quantity.

In [12], Braasch, Rein, and Vukadinović consider the case where μ can be factorized as

$$\mu(E,I) = \gamma(E-\alpha)\,\nu(I) \quad \text{for all} \ \ (E,I) \in \mathbb{R}^2\ ,$$

where γ is compactly supported and $\alpha \in \mathbb{R}$. If γ satisfies (H3) and (H4) and if ν is a C^1 uniformly positive function, our previous results can easily be extended. In this section, we are going to consider general steady states corresponding to functions μ which cannot be factorized in terms of two functions γ and ν (such an extension has already been considered by Guo and Rein in [32] for gravitational systems) or which do not necessarily have a compact support in E.

In order to obtain the existence of these stationary solutions, we have to assume the following hypotheses on μ and ϕ_e, which are generalizations of (H3) and (H4) of section 2.

(H3′) *Let $\sigma:\ \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R}$ be such that $\frac{\partial\sigma}{\partial s}(s,I) = -\mu^{-1}(s,I)$ and assume that for any fixed $I \in \mathbb{R}$, $\sigma(.,I)$ has a $C^0({\mathbb{R}_0}^+) \cap C^2(\mathbb{R}^+)$ regularity and is bounded from below, strictly convex, and such that $\lim_{s\to+\infty} \sigma(s,I)/s = +\infty$. Here μ^{-1} is the generalized inverse of $s \mapsto \mu(s,I)$ for fixed I.*

(H4′) *The external potential $\phi_e : \mathbb{R}^3 \to \mathbb{R}$ is a measurable bounded-from-below function such that $\lim_{|x|\to+\infty} \phi_e(x) = +\infty$ and*

$$x \mapsto \int_{\mathbb{R}^3} \mu\left(\frac{1}{2}|v|^2 + \phi_e(x), I(x,v)\right)\ dv$$

belongs to $L^1 \cap L^\infty(\mathbb{R}^3)$.

The stationary solution $f_{\infty,\sigma}$ is characterized as the unique nonnegative critical point of a strictly convex coercive functional K_σ, with

$$K_\sigma[f] = \int_{\mathbb{R}^6} \left[\sigma(f,I) + \left(\frac{1}{2}|v|^2 + \phi_e(x)\right) f\right]\ d(x,v) + \frac{1}{2}\int_{\mathbb{R}^3} |\nabla\phi[f]|^2\ dx\ ,$$

under the constraint $\int_{\mathbb{R}^6} f_{\infty,\sigma}\ d(x,v) = M$ for some given $M > 0$. As in section 2, α_M in (6.1) is the Lagrange multiplier associated to the constraint on the mass and is uniquely determined by the condition $\int_{\mathbb{R}^6} f_{\infty,\sigma}\ d(x,v) = M$. To σ we associate a relative entropy functional defined by

$$\begin{aligned}\Sigma_\sigma[f|f_{\infty,\sigma}] &:= K_\sigma[f] - K_\sigma[f_{\infty,\sigma}]\\ &= \int_{\mathbb{R}^6} \left[\sigma(f,I) - \sigma_\infty - \frac{\partial\sigma_\infty}{\partial s}\,(f - f_{\infty,\sigma})\right] d(x,v) + \frac{1}{2}\int_{\mathbb{R}^3} |\nabla_x(\phi[f] - \phi_{\infty,\sigma})|^2 dx\end{aligned}$$

with $\sigma_\infty = \sigma(f_{\infty,\sigma}, I)$ and $\phi_{\infty,\sigma} = \phi[f_{\infty,\sigma}]$.

If there exists a function $A_\sigma(I) > 0$ such that $\frac{\partial^2\sigma}{\partial s^2}(s,I) \ge A_\sigma(I)$ for any $(s,I) \in \mathbb{R}_0^+ \times \mathbb{R}$, by Taylor expansion it follows that

$$\Sigma_\sigma[f|f_{\infty,\sigma}] \ge \int_{\mathbb{R}^6} A_\sigma(I)|f - f_{\infty,\sigma}|^2\ d(x,v)\ ,$$

which proves a weighted L^2-stability result. Exactly as before, we can use a cut-off argument and get a generalization of Theorem 4.3.

Let $E_\sigma(x,v) := \frac{1}{2}|v|^2 + \phi_{\infty,\sigma}(x) + \phi_e(x)$ and $E_{min} := \inf\{E_\sigma(x,v) \ : \ (x,v) \in \mathbb{R}^6\}$, which is finite by assumption (H4′). With evident notation, $f_{\infty,\sigma} = \mu(E_\sigma(\cdot) - \alpha_M, I(\cdot,\cdot))$. Take $\bar{s}(I) = \mu(E_{min} - \alpha_M, I)$ and define for any $I \in \mathbb{R}$

$$\tau_\sigma(s,I) := \begin{cases} \sigma(s,I) & \text{if } s \in [0,\bar{s}(I)] \ , \\ \psi(s,I) & \text{if } s \in (\bar{s}(I),+\infty) \end{cases} \tag{6.2}$$

with $\psi(s,I) = \frac{\sigma''(\bar{s},I)}{\sigma_2''(\bar{s})}\sigma_2(s) + (\sigma'(\bar{s},I) - \frac{\sigma''(\bar{s},I)}{\sigma_2''(\bar{s})}\,\sigma_2'(\bar{s}))(s-\bar{s}) + \sigma(\bar{s},I) - \frac{\sigma''(\bar{s},I)}{\sigma_2''(\bar{s})}\,\sigma_2(\bar{s})$, $\bar{s} = \bar{s}(I)$, and $\sigma_2(s) = s^2$. With the truncated Lyapunov functional $\Sigma_{\tau_\sigma}[f|f_{\infty,\sigma}] = K_{\tau_\sigma}[f] - K_{\tau_\sigma}[f_{\infty,\sigma}]$, we immediately get the following variant of Theorem 4.3.

THEOREM 6.1. *Let I be a function in $C^1(\mathbb{R}^6)$ and assume that ϕ_e, μ verify* (H3′)–(H4′). *Assume, moreover, that*

$$B_\sigma(I) = \inf\left\{ s \in [E_{min} - \alpha_M, \mu^{-1}(0,I)] \ : \ \frac{\partial^2\sigma}{\partial s^2}(s,I) \right\} > 0 \quad \textit{for any } I \in \mathbb{R} \ .$$

Let f_0 be a nonnegative function in $L^1(\mathbb{R}^6) \cap L^2(\mathbb{R}^6, B_\sigma(I(x,v))\,d(x,v))$ such that $(x,v) \mapsto \sigma(f_0(x,v), I(x,v))$ belongs to $L^1(\mathbb{R}^6)$ and consider a weak (resp., renormalized) solution of the Vlasov–Poisson system with initial data f_0 satisfying $(\mathcal{W})$ (resp., $(\mathcal{R})$). Then for any $t \geq 0$

$$\Sigma_{\tau_\sigma}[f_0|f_{\infty,\sigma}] \geq \Sigma_{\tau_\sigma}[f(t)|f_{\infty,\sigma}] \geq \int_{\mathbb{R}^6} B_\sigma(I(x,v))\,|f(t,x,v) - f_{\infty,\sigma}(x,v)|^2 \ d(x,v) \ .$$

Weighted L^q estimates can also be established if one replaces σ_2 by σ_q in (6.2), under the condition that $\inf\{s \in [E_{min} - \alpha, \mu^{-1}(0,I)] \ : \ s^{2-q}\frac{\partial^2\sigma}{\partial s^2}(s,I)\} > 0$ for any $I \in \mathbb{R}$.

Remark 6.2. Equation (1.1) is a special case (parabolic-band approximation) of the Vlasov–Poisson system for semiconductors

$$\frac{\partial f}{\partial t} + v(p)\cdot\nabla_x f + F(t,x)\cdot\nabla_p f = 0$$

on $\mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3$, with $v(p) = \nabla_p \epsilon(p)$. If we assume that ϵ is a nonnegative C^1 function such that $e^{-\epsilon(p)} \in L^1(\mathbb{R}^3)$, then abusing the same notations as for (1.1) (which corresponds to the special case $\epsilon(p) = \frac{1}{2}p^2$), one can, for instance, prove that there exists a Maxwellian-type stationary solution given by

$$m(x,p) \ = \ M\,\frac{e^{-\epsilon(p) - q\,(\phi(x) + \phi_e(x))}}{\int_{\mathbb{R}^6} e^{-\epsilon - q\,(\phi+\phi_e)}\ d(x,p)} \ ,$$

where ϕ is given by (1.3) with $\rho(f)(t,x) = \int_{\mathbb{R}^3} f(t,x,p)\ dp$. Nonlinear stability results for m and more general stationary states can be easily obtained using the previous ideas. Realistic models include collisions, which usually determine a special class of stationary solutions (and the appropriate Lyapunov functional is then decreasing even for classical solutions). We refer to [37, 6, 7, 17, 19] for more details on this subject.

7. Appendix: A convexity property of L^1 functions. Let f_0 be a nonnegative function in $L^1(\Omega)$ for some (not necessarily bounded) domain Ω in $\mathbb{R}^d$, $d \geq 1$. It is straightforward to check that $\sigma(f_0) \in L^1(\Omega)$ if σ is a C^2 convex function on $\mathbb{R}^+$

such that $s \mapsto \sigma(s)/s$ is bounded (consider, for example, $\sigma(s) = 2s + e^{-s} - 1$). The result of Proposition 2.1, which is a special case of the following Proposition, is much stronger.

PROPOSITION 7.1. *Let $(E, d\mu)$ be a measurable space. For any nonnegative function f_0 in $L^1(E, d\mu)$, there exists a nonnegative strictly convex function σ of class C^2 such that $\lim_{s\to+\infty} \sigma(s)/s = +\infty$ and $\sigma(f_0) \in L^1(E, d\mu)$.*

This result is more or less standard. For completeness, we are going to give a proof which is based on the following elementary lemma.

LEMMA 7.2. *Consider a sequence $\{\alpha_n\}$ with $\alpha_n > 0$ for any n and $\sum \alpha_n < \infty$. Then there exists an increasing sequence $\{\beta_n\}$ with $\beta_n > 0$ for any $n \in \mathbb{N}$, and $\lim_{n\to\infty} \beta_n = +\infty$ such that $\sum \alpha_n\beta_n < \infty$.*

Proof of Lemma 7.2. We prove this result by an explicit construction of β_n. Let $\epsilon_n = \sum_{m\geq n} \alpha_m$ and take $\beta_n = \frac{1}{2\sqrt{\epsilon_n}}$:

$$\alpha_n\beta_n = (\epsilon_n - \epsilon_{n+1}) \frac{1}{2\sqrt{\epsilon_n}} \leq \sqrt{\epsilon_n} - \sqrt{\epsilon_{n+1}} \, ,$$

which immediately gives $\sum_{m\geq n} \alpha_m\beta_m \leq \sqrt{\epsilon_n}$. □

Proof of Proposition 7.1. Let $\alpha_n = \int_{n\leq f_0<n+1} f_0 \, d\mu$ and take β_n given by Lemma 7.2. One can find a convex function σ with $s \mapsto \sigma(s)/s$ nondecreasing such that $\sigma(n + 1) = (n + 1)\beta_n$. Thus

$$\int_{n\leq f_0<n+1} \sigma(f_0) \, d\mu \leq \int_{n\leq f_0<n+1} f_0 \, d\mu \cdot \frac{\sigma(n+1)}{n+1} = \alpha_n\beta_n \, ,$$

which ends the proof. □

Remark 7.3. From Proposition 7.1, it is clear that there is no optimal convex function σ corresponding to a given initial data f_0 (reapply the Proposition to $\sigma(f_0)$). To any σ, one can, however, associate a function γ. Is there an optimal condition on the growth of ϕ_e so that both the stationary solution and the relative entropy are well defined? This would indeed define a notion of *confinement* which would depend only on f_0. On the other hand, if the growth condition is not satisfied, is it possible to give some dispersion estimate (as in the case $\phi_e \equiv 0$, or $(x - x_0) \cdot \nabla\phi_e \geq 0$ for some given $x_0 \in \mathbb{R}^3$)?

Acknowledgment. The authors thank Bernt Wennberg for a comment which led us to include the results stated in the appendix.

REFERENCES

[1] A. ARNOLD, P. MARKOWICH, G. TOSCANI, AND A. UNTERREITER, *On generalized Csiszár-Kullback inequalities*, Monatsh. Math., 131 (2000), pp. 235–253.

[2] A. ARNOLD, P. MARKOWICH, G. TOSCANI, AND A. UNTERREITER, *On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations*, Comm. Partial Differential Equations, 26 (2001), pp. 43–100.

[3] A.A. ARSEN'EV, *Global existence of a weak solution of Vlasov's system of equations*, Soviet Math. Dokl., 14 (1973), pp. 1763–1765.

[4] J. BATT, P.J. MORRISON, AND G. REIN, *A rigorous stability result for the Vlasov-Poisson system in three dimensions*, Ann. Mat. Pura Appl. (4), 164 (1993), pp. 133–154.

[5] J. BATT, P.J. MORRISON, AND G. REIN, *Linear stability of stationary solutions of the Vlasov-Poisson system in three dimensions*, Arch. Ration. Mech. Anal., 130 (1995), pp. 163–182.

[6] N. BEN ABDALLAH AND J. DOLBEAULT, *Entropies relatives pour le système de Vlasov-Poisson dans des domaines bornés (Relative entropies for the Vlasov-Poisson system in bounded domains)*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 867–872.

[7] N. Ben Abdallah and J. Dolbeault, *Relative entropies for kinetic equations in bounded domains (irreversibility, stationary solutions, uniqueness)*, Arch. Ration. Mech. Anal., to appear.

[8] P. Biler and J. Dolbeault, *Long time behavior of solutions to Nernst-Planck and Debye-Hückel drift-diffusion systems*, Ann. Henri Poincaré, 1 (2000), pp. 461–472.

[9] P. Biler, J. Dolbeault, and P. Markowich, *Large time asymptotics of nonlinear drift-diffusion systems with Poisson coupling*, Transport Theory Statist. Phys., 30 (2001), pp. 521–536.

[10] L.L. Bonilla, J.A. Carrillo, and J. Soler, *Asymptotic behavior of an initial-boundary value problem for the Vlasov–Poisson–Fokker–Planck system*, SIAM J. Appl. Math., 57 (1997), pp. 1343–1372.

[11] F. Bouchut and J. Dolbeault, *On long time asymptotics of the Vlasov-Fokker-Planck equation and of the Vlasov-Poisson-Fokker-Planck system with Coulombic and Newtonian potentials*, Differential Integral Equations, 8 (1995), pp. 487–514.

[12] P. Braasch, G. Rein, and J. Vukadinović, *Nonlinear stability of stationary plasmas—an extension of the energy-Casimir method*, SIAM J. Appl. Math., 59 (1998), pp. 831–844.

[13] J.A. Carrillo and G. Toscani, *Asymptotic L^1-decay of solutions of the porous medium equation to self-similarity*, Indiana Univ. Math. J., 49 (2000), pp. 113–141.

[14] J.A. Carrillo, A. Jungel, P. Markowich, G. Toscani, and A. Unterreiter, *Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities*, Monatsh. Math., 133 (2001), pp. 1–82.

[15] F. Castella, *Effets dispersifs dans les équations de Schrödinger et de Vlasov*, Séminaire sur les Equations aux Dérivées Partielles, 1997-1998, Exp. No. XXIV, 14 pp., École Polytechnique, Palaiseau, 1998.

[16] F. Castella, *Propagation of space moments in the Vlasov-Poisson equation and further results*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 16 (1999), pp. 503–533.

[17] C. Cercignani, I.M. Gamba, and C.L. Levermore, *A high field approximation to a Boltzmann-Poisson system in bounded domains*, Appl. Math Lett., 4 (1997), pp. 111–118.

[18] P. Degond, *Spectral theory of the linearized Vlasov-Poisson equation*, Trans. Amer. Math. Soc., 294 (1986), pp. 435–453.

[19] P. Degond, F. Poupaud, B. Niclot, and F. Guyot, *Semiconductor modelling via the Boltzmann equation*, in Computational Aspect of VLSI Design with an Emphasis on Semiconductor Device Simulation, Lectures in Appl. Math. 25, 1990, pp. 51–73.

[20] M. Del Pino and J. Dolbeault, *Generalized Sobolev Inequalities and Asymptotic Behaviour in Fast Diffusion and Porous Media Problems*, Preprint Ceremade 9905, Université Paris IX, 1999, pp. 1–45.

[21] M. Del Pino and J. Dolbeault, *Best constants for Gagliardo-Nirenberg inequalities and application to nonlinear diffusions*, J. Math. Pures Appl. (9), to appear.

[22] M. Del Pino and J. Dolbeault, *Nonlinear diffusions and optimal constants in Sobolev type inequalities: Asymptotic behaviour of equations involving the p-Laplacian*, C. R. Math. Acad. Sci. Paris, 334 (2002), pp. 365–370.

[23] M. Del Pino and J. Dolbeault, *Asymptotic Behaviour of Nonlinear Diffusions*, Preprint Ceremade 0127, Université Paris IX, 2001, pp. 1–8.

[24] L. Desvillettes and C. Villani, *On the trend to global equilibrium in spatially inhomogeneous entropy-dissipating systems. Part* I: *The linear Fokker-Planck equation*, Comm. Pure Appl. Math., 54 (2001), pp. 1–42.

[25] R.-J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[26] R.-J. DiPerna and P.-L. Lions, *Solutions globales d'équations du type Vlasov–Poisson*, C. R. Acad. Sci. Paris Sér. I Math., 307 (1988), pp. 655–658.

[27] J. Dolbeault, *Free energy and solutions of the Vlasov-Poisson-Fokker-Planck system: External potential and confinement (large time behavior and steady states)*, J. Math. Pures Appl. (9), 78 (1999), pp. 121–157.

[28] R.T. Glassey, *The Cauchy Problem in Kinetic Theory*, SIAM, Philadelphia, 1996.

[29] Y. Guo, *Stable magnetic equilibria in collisionless plasmas*, Comm. Pure Appl. Math., 50 (1997), pp. 891–933.

[30] Y. Guo, *Variational method for stable polytropic galaxies*, Arch. Ration. Mech. Anal., 150 (1999), pp. 209–224.

[31] Y. Guo and G. Rein, *Existence and stability of Camm type steady states in galactic dynamics*, Indiana Univ. Math. J., 48 (1999), pp. 1237–1255.

[32] Y. Guo and G. Rein, *Stable steady states in stellar dynamics*, Arch. Ration. Mech. Anal., 147 (1999), pp. 225–243.

[33] E. Hörst, *Global strong solutions of Vlasov's equations, necessary and sufficient conditions for their existence*, in Partial Differential Equations, Banach Center Publ. 19, Warsaw, 1987, pp. 143–153.
[34] E. Hörst and R. Hunze, *Weak solutions of the initial value problem for the unmodified nonlinear Vlasov equation*, Math. Methods Appl. Sci., 6 (1984), pp. 262–279.
[35] P.-L. Lions, *Sur les équations différentielles ordinaires et les équations de transport*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 833–838.
[36] P.L. Lions and B. Perthame, *Propagation of moments and regularity for the Vlasov-Poisson system*, Invent. Math., 105 (1991), pp. 415–430.
[37] P.A. Markowich, C.A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, New York, 1990.
[38] S. Mischler, *On the trace problem for solutions of the Vlasov equation*, Comm. Partial Differential Equations, 25 (2000), pp. 1415–1443.
[39] F. Otto, *The geometry of dissipative evolution equations: The porous medium equation*, Comm. Partial Differential Equations, 26 (2001), pp. 101–174.
[40] B. Perthame, *Time decay, propagation of low moments and dispersive effects for kinetic equations*, Comm. Partial Differential Equations, 21 (1996), pp. 659–686.
[41] K. Pfaffelmoser, *Global classical solutions of the Vlasov-Poisson system in three dimensions for general initial data*, J. Differential Equations, 95 (1992), pp. 281–303.
[42] G. Rein, *Non-linear stability for the Vlasov-Poisson system—the energy-Casimir method*, Math. Methods Appl. Sci., 17 (1994), pp. 1129–1140.
[43] J. Schaeffer, *Global existence of smooth solutions to the Vlasov-Poisson system in three dimensions*, Comm. Partial Differential Equations, 16 (1991), pp. 1313–1335.
[44] G. Wolansky, *On nonlinear stability of polytropic galaxies*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 16 (1999), pp. 15–48.

SIAM J. MATH. ANAL.
Vol. 34, No. 2, pp. 495–509

LOCAL STRESS REGULARITY IN SCALAR NONCONVEX VARIATIONAL PROBLEMS*

CARSTEN CARSTENSEN† AND STEFAN MÜLLER‡

Abstract. In light of applications to relaxed problems in the calculus of variations, this paper addresses convex but not necessarily strictly convex minimization problems. A class of energy functionals is described for which any stress field σ in $L^q(\Omega)$ with $\operatorname{div}\sigma$ in $W^{1,p'}(\Omega)$ belongs to $W^{1,q}_{loc}(\Omega)$. The condition on $\operatorname{div}\sigma$ holds, for example, for solutions of the Euler–Lagrange equations involving additional lower-order terms. Applications include the scalar double-well potential, an optimal design problem, a vectorial double-well problem in a compatible case, and Hencky elastoplasticity with hardening. If the energy density depends only on the modulus of the gradient, we also show regularity up to the boundary.

Key words. nonconvex minimization, regularization, relaxed problem, stress regularity

AMS subject classifications. 49N60, 74B15, 74G40, 74N15, 35D10, 35J70

PII. S0036141001396436

1. Introduction.

Consider a volume term $f \in L^q_{loc}(\Omega)$, Dirichlet data $u_0 \in W^{1,p}(\Omega)$, and a nonvoid, closed, convex set $\mathcal{A}$ of admissible displacements, which satisfies $u_0 + W^{1,p}_0(\Omega) \subseteq \mathcal{A} \subseteq W^{1,p}(\Omega)$. The problem

$$\text{minimize } E(u) := \int_\Omega W(Du)\,dx - \int_\Omega f\,u\,dx \quad \text{among } u \in \mathcal{A} \tag{1.1}$$

may fail to have a solution in $\mathcal{A}$. Typically, infimizing sequences exist and are bounded in the seminorm of $W^{1,p}(\Omega)$ and weakly convergent towards some u in $\mathcal{A}$. The limit u, however, may fail to minimize the energy E since the functional $E : \mathcal{A} \to \mathbb{R}$ is not (sequentially) weakly lower semicontinuous owing to its nonconvexity.

Nevertheless, u describes the macroscopic, space-averaged state and is therefore of interest. Relaxation results in the calculus of variations show that u can be computed as a solution of the relaxed problem

$$\text{minimize } RE(u) := \int_\Omega \varphi(Du)\,dx - \int_\Omega f\,u\,dx \quad \text{among } u \in \mathcal{A}. \tag{1.2}$$

In the general case φ is the quasi-convexification of W [Dac89, Rou97]. The arguments of this paper are essentially restricted to the situation where φ is the convex envelope of W.

It was observed in [Fri94, Cel93a, Cel93b, CP97b] for scalar problems and recently in [BKK00] in the general case that the stress fields $\sigma_j := DW(Du_j)$ of an infimizing sequence u_j converge in a weak sense. The limit σ is given as the stress of a relaxed

*Received by the editors October 15, 2001; accepted for publication (in revised form) May 8, 2002; published electronically December 11, 2002.
http://www.siam.org/journals/sima/34-2/39643.html

†Institute for Applied Mathematics and Numerical Analysis, Vienna University of Technology, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria (Carsten.Carstensen@tuwien.ac.at). This author's research was supported by the Max Planck Institute for Mathematics in the Sciences in Leipzig, Germany, and the German Research Foundation through the DFG-Schwerpunktprogramm Multi-Scale Problems (SPP 1095).

‡Max Planck Institute for Mathematics in the Sciences, Inselstr. 22-26, D-04103 Leipzig, Germany (Stefan.Mueller@mis.mpg.de).

energy φ, i.e., $\sigma = D\varphi(Du)$. Hence the stress field associated with (1.1) can be computed from (1.2); for the regularity of σ it suffices to study (1.2).

This paper establishes local regularity of the stress variable σ under minimal conditions on u. We consider a class of convex (but not necessarily strictly convex) C^1 functions φ with

$$|D\varphi(A) - D\varphi(B)|^2 \le c_1\,(1 + |A|^s + |B|^s) \times (D\varphi(A) - D\varphi(B)) : (A - B) \tag{1.3}$$

for all $A, B \in \mathbb{M}^{m\times n}$ ($\mathbb{M}^{m\times n}$ denotes the real $m \times n$ matrices) and a multiplicative constant c_1. Note that (1.3) implies convexity of φ but not strict convexity.

Theorem 2.1 of section 2 asserts that the monotonicity condition (1.3) and the identity $\operatorname{div} D\varphi(Du) = f$ in $W^{1,q}(\Omega)$, which is the Euler–Lagrange equation corresponding to (1.2), together yield $\sigma = D\varphi(Du)$ in $W^{1,q}_{loc}(\Omega)$. Examples include the scalar two-well potential (see section 3) and a relaxed energy density of an optimal design problem (see section 4).

A symmetric variant of (1.1)–(1.2), where $n = m$ and where Du is replaced by the symmetric part $\varepsilon(u) := \operatorname{sym} Du$, is given by

$$\text{minimize } RE(u) := \int_\Omega \varphi(\varepsilon(u))\,dx - \int_\Omega f\,u\,dx \quad \text{among } u \in \mathcal{A} \tag{1.4}$$

and is discussed in section 5. Emphasis is put on the robustness of the stress estimate as $\lambda \to \infty$, where λ is the Lamé constant related to volume changes. Applications to Hencky elastoplasticity and a vector two-well example in sections 6 and 7, respectively, conclude this paper.

Throughout this paper $\mathbb{M}^{m\times n}$ denotes the real $m \times n$ matrices endowed with the Euclidean scalar product $A : B := \sum_{j=1}^m \sum_{k=1}^n A_{jk}B_{jk}$ and the induced (Frobenius matrix) norm $|\cdot|$, $|A| := (A : A)^{1/2}$. We use standard notation for Sobolev and Lebesgue spaces and their norms and seminorms.

2. Abstract stress regularity result. Let Ω be an open set in $\mathbb{R}^n$, let $\varphi : \mathbb{M}^{m\times n} \to \mathbb{R}$ be a C^1 function, and let $D\varphi$ be its derivative. Suppose that there exist constants $1 < p < \infty$, $1 < r < \infty$, $0 \le s < \infty$, and $0 < c_2$ such that, for all $A, B \in \mathbb{M}^{m\times n}$,

$$|D\varphi(A) - D\varphi(B)|^r \le c_2\,(1 + |A|^s + |B|^s) \times (D\varphi(A) - D\varphi(B)) : (A - B). \tag{2.1}$$

THEOREM 2.1. *Assume furthermore that*

$$u \in W^{1,p}(\Omega;\mathbb{R}^m) \quad \textit{and} \quad \sigma := D\varphi(Du)$$

satisfy

$$\sigma \in L^q_{loc}(\Omega;\mathbb{M}^{m\times n}) \quad \textit{and} \quad \operatorname{div}\sigma \in W^{1,p'}_{loc}(\Omega;\mathbb{R}^m)$$

for $p' := p/(p-1)$ *and* $q := r/(1 + s/p)$. *Suppose* $p' \le q$ *and* $r \le 2$. *Then*

$$\sigma \in W^{1,q}_{loc}(\Omega;\mathbb{M}^{m\times n}).$$

Remark 2.1. (a) The point is that (2.1) implies that φ is convex; nonetheless, φ need not to be strictly convex since the lower bound is in terms of stress differences but not in terms of $|A - B|$.

(b) The assumptions on u can be localized to $u \in W^{1,p}_{loc}(\Omega;\, \mathbb{R}^n)$ by applying the result to subsets of Ω.

Proof. Let ω be an open, bounded set which is compactly contained in Ω, i.e., $\omega \subset \overline{\omega} \subset \Omega$. Fix $\eta \in \mathcal{D}(\omega)$ and a direction $M \in \mathbb{M}^{m\times n}$ with $|M| = 1$. Set $\alpha := 1/(r-1)$ and $\beta := r/(r-1)$. For $0 < h < h_0 := \operatorname{dist}(\operatorname{supp}\eta; \partial\omega)$ consider the difference quotients

$$\begin{aligned} \tau(x) &:= (\sigma(x+hM) - \sigma(x))/h, \\ e(x) &:= (u(x+hM) - u(x))/h, \\ \delta(x) &:= De(x). \end{aligned}$$

A standard argument in the approximation of weak derivatives by difference quotients shows that

$$\|e\|_p := \|e\|_{L^p(\omega)} \le c_3\, \|u\|_{W^{1,p}(\Omega)}. \tag{2.2}$$

Here and throughout the proof $\|\cdot\|_t := \|\cdot\|_{L^t(\omega)}$ denotes the $L^t(\omega)$-norm with respect to the subdomain ω of Ω.

Since $u \in W^{1,p}(\Omega)$ the expression $\|e\|_{L^p(\omega)}$ is bounded uniformly in h. A careful use of Hölder's inequality in combination with $q' \le p$,

$$\operatorname{div}\sigma \in W^{1,p'}(\omega; \mathbb{M}^{m\times n}) \quad \text{and} \quad \sigma \in L^{p'}(\omega; \mathbb{M}^{m\times n}),$$

yields the uniform bound

$$\|\, \varrho^{q/r}\, \|^{r/q}_{1+p/s} + \|e\|_p + \|e\|_{q'} + \|\eta^\beta \operatorname{div}\tau\|_{p'} + \|\,\eta\,\|_{W^{1,\infty}(\Omega)} \le c_4, \tag{2.3}$$

where $\varrho(x) := 1 + |Du(x)|^s + |Du(x+hM)|^s$.

To verify the assertion, we have to bound $|\tau|_{L^q(K)}$ uniformly in h for each compact set $K \subset \Omega$ (below K is a compact subset of the interior of $\operatorname{supp}\eta$).

Applying (2.1) with $A := Du(x+hM)$ and $B := Du(x)$, we obtain

$$|\tau|^r \le c_2\, h^{2-r}\, \varrho\, \tau : \delta \quad \text{a.e. in } \omega. \tag{2.4}$$

Raising (2.4) to the power q/r, multiplying with $\eta^{\alpha q}$, and finally integrating the result over Ω, we infer that

$$\|\eta^\alpha \tau\|_q^q \le c_2^{q/r}\, h^{q(2-r)/r} \int_\Omega \eta^{\alpha q}\, \varrho^{q/r}\, (\tau:\delta)^{q/r} dx. \tag{2.5}$$

Applying Hölder's inequality (with r/q and $(r/q)' = 1 + p/s$), raising the result to the power r/q, and using the fact that $0 \le \tau : \delta$ and $\alpha r = \beta$, we derive

$$\|\eta^\alpha \tau\|_q^r \le c_2\, h^{2-r}\, \|\, \varrho^{q/r}\, \|^{r/q}_{1+p/s} \int_\Omega \eta^\beta\, \tau : \delta\, dx. \tag{2.6}$$

Since $\delta = De$ on ω and $h \le h_0$, an integration by parts proves that

$$\begin{aligned} \int_\Omega \eta^\beta\, \tau : \delta\, dx &= -\int_\Omega e \cdot \operatorname{div}(\eta^\beta \tau) dx \\ &\le \beta \|\,\eta\,\|_{1,\infty}\, \|\, \eta^{\beta-1}\, \tau\, e\,\|_1 + \|\, e\,\|_p \|\, \eta^\beta \operatorname{div}\tau\,\|_{p'}. \end{aligned} \tag{2.7}$$

Hölder's inequality and the relation $\beta - 1 = \alpha$ lead to

$$\| \eta^{\beta-1} \tau e \|_1 \leq \| e \|_{q'} \| \eta^{\alpha} \tau \|_q. \tag{2.8}$$

The combination of (2.6)–(2.8) with (2.3) and the hypothesis $r \leq 2$ proves that

$$\|\eta^{\alpha}\tau\|_q^r \leq c_2 c_4^3 h_0^{2-r}(1 + \| \eta^{\alpha}\tau \|_q). \tag{2.9}$$

With Young's inequality $ab \leq (ac)^r/r + (b/c)^{r'}/r'$ for positive a, b, c we deduce from (2.9) and the assumptions $r \leq 2$ and $q > 1$ that $\|\eta^{\alpha}\tau\|_q$ is bounded uniformly in h. Hence,

$$\limsup_{h\to 0} \|\eta^{\alpha}\tau\|_{L^q(\Omega)} < \infty \quad \text{for all } \eta \in \mathcal{D}(\Omega).$$

The proof is finished. □

To illustrate the growth condition in (2.1) we consider the simple example of the p-Laplace equation.

Example 2.1. Let $m = 1$, let $2 \leq p < \infty$, and consider $\varphi(F) := |F|^p/p$ for $F \in \mathbb{R}^n$. Then $D\varphi(F) = |F|^{p-2} F$, and for fixed $B \in \mathbb{R}^n$ and $A \in \mathbb{R}^n$ with $|A| \to \infty$, we have

$$\frac{|D\varphi(A) - D\varphi(B)|^2}{(D\varphi(A) - D\varphi(B) \cdot (A - B)} \approx |D\varphi(A)|/|A| = |A|^{p-2}.$$

Indeed, it is known (e.g., by a combination of Lemmas 2.1 to 2.3 in [CK01]) that for any $A, B \in \mathbb{R}^n$,

$$\frac{|D\varphi(A) - D\varphi(B)|^2}{(D\varphi(A) - D\varphi(B) \cdot (A - B)} \leq (1 + \max\{1, p-2\}^2)(|A|^{p-2} + |B|^{p-2}).$$

We therefore obtain, as a corollary of Theorem 2.1, local regularity of the stress field, i.e., $\sigma := D\varphi(Du) \in W^{1,p'}_{loc}(\Omega; \mathbb{R}^n)$ for a minimizer $u \in W^{1,p}(\Omega)$ of (1.2) with $f \in W^{1,p'}_{loc}(\Omega)$.

3. An application to the scalar two-well problem. This section concerns the scalar double-well problem, where

$$W : \mathbb{R}^n \to \mathbb{R}, \quad F \mapsto |F - F_1|^2 \, |F - F_2|^2 \tag{3.1}$$

and where the energy wells $F_1, F_2 \in \mathbb{R}^n$, $F_1 \neq F_2$, are given. The scalar problem (1.1) with (3.1) (for $m = 1$) can be deduced from the Ericksen–James energy density in an antiplane shear model; the version for $n = 1$, due to Bolza [Bol06]) (cf. also [You69]), is the model example in nonconvex minimization.

PROPOSITION 3.1 (see [CP97b]). *Let $a := (F_2 - F_1)/2$ and $b := (F_1 + F_2)/2$. The convex envelope φ of W given by* (3.1) *is*

$$\varphi(F) := \max\{|F - b|^2 - |a|^2, 0\}^2 + 4(|a|^2 \, |F - b|^2 - [a \cdot (F - b)]^2)$$

and satisfies (2.1) *with $r = 2$, $s = 2$, and $c_2 = 4 \max\{2, |F_1 - F_2|^2\}$.*

COROLLARY 3.2. *Adopt the notation of Proposition* 3.1 *and let u be a minimizer of* (1.2) *with $f \in W^{1,4/3}$. Then $\sigma := D\varphi(Du) \in W^{1,4/3}_{loc}(\Omega; \mathbb{R}^n)$.*

Proof. The assertion follows from Theorem 2.1 and Proposition 3.1 since the Euler–Lagrange equations of the minimization problem (1.2) imply that $-\operatorname{div}\sigma = f \in W^{1,4/3}(\Omega)$. □

Remark 3.1. Further estimates in [CP97b] allow one to control other quantities. In particular,

$$\max\{0, |B - Du|^2 - |A|^2\} \in H^1_{loc}(\Omega) \quad \text{and} \quad M \cdot Du \in H^1_{loc}(\Omega)$$

for all directions M perpendicular to A.

4. An application to an optimal design problem. The relaxed model for an optimal design problem derived in [GKR86] has the form (1.2), where $\varphi(F) = \psi(|F|)$. Given positive parameters $0 < t_1 < t_2$ and $0 < \mu_2 < \mu_1$ with $t_1\mu_1 = t_2\mu_2$, the C^1 function $\psi : [0,\infty) \to [0,\infty)$ is defined by $\psi(0) = 0$ and

$$\psi'(t) := \begin{cases} \mu_1\, t & \text{if } 0 \le t \le t_1, \\ t_1\mu_1 = t_2\mu_2 & \text{if } t_1 \le t \le t_2, \\ \mu_2\, t & \text{if } t_2 \le t. \end{cases}$$

PROPOSITION 4.1 (see [CP97b]). *The function $\varphi(F) = \psi(|F|)$ satisfies (2.1) with $r = 2$, $s = 0$, and $c_2 = 1/\mu_1$.*

Therefore Theorem 2.1 yields local stress regularity for minimizers of (1.2) when $f \in H^1_{loc}(\Omega)$.

COROLLARY 4.2. *Adopt the notation of Proposition 4.1 and let u be a minimizer of (1.2) in $\mathcal{A} := H^1_0(\Omega)$. Then $\sigma := D\varphi(Du) \in W^{1,2}_{loc}(\Omega;\mathbb{R}^n)$.*

Global regularity of the variable $u \in C^\alpha(\overline{\Omega}) \cap W^{1,\infty}(\Omega)$ and of the level sets has been analyzed in [KSW91].

The rest of this section is devoted to establishing stress regularity up to the boundary.

THEOREM 4.3. *Suppose that $f \in W^{1,2}_0(\Omega)$ and that Ω is a $C^{2,1}$ domain. If u is a minimizer of (1.2), then $\sigma := D\varphi(Du) \in W^{1,2}(\Omega;\mathbb{R}^n)$.*

The remaining part of this section is devoted to a proof of Theorem 4.3 via a local reflection argument. Owing to the local regularity of Corollary 4.2, it remains to prove $\sigma \in W^{1,2}(\Omega \cap B(x_0,\delta);\mathbb{R}^n)$ for each point x_0 on the boundary $\partial\Omega$ and some small $\delta > 0$. Without loss of generality, we suppose $x_0 = 0$ and that the Cartesian coordinate system at hand directly allows a $C^{2,1}$ parameterization.

DEFINITION 4.1. *Let $\chi : B'_0 \to \mathbb{R}$ be a (scalar) $C^{2+\alpha}$ function, where $B_0 := B(0,\delta_0) \subset \mathbb{R}^n$ and $B'_0 := \{x' \in \mathbb{R}^{n-1} : |x'| < \delta_0\} \subset \mathbb{R}^{n-1}$ denote the δ_0-ball around $x_0 = 0$ in n and $(n-1)$ dimensions, respectively. Suppose that χ parameterizes the boundary $\Gamma := \partial\Omega$ near $x_0 = 0$, i.e.,*

$$\begin{aligned} \Gamma \cap B_0 &= \{(x', \chi(x')) \in B_0 : x' \in B'_0\}, \\ \Omega \cap B_0 &= \{(x', x_n) \in B_0 : x' \in B', x_n > \chi(x')\}, \\ B_0 \backslash \overline{\Omega} &= \{(x', x_n) \in B_0 : x' \in B', x_n < \chi(x')\}. \end{aligned}$$

Let ν be the unit normal vector on Γ and set

$$\Psi(x) := (x', \chi(x')) - x_n\, \nu(x', \chi(x')) \quad \text{for all } x =: (x', x_n) \in B_0 \subset B'_0 \times \mathbb{R}.$$

LEMMA 4.4. *The pull-back metric $g := D\Psi^T D\Psi : B_0 \to \mathbb{M}^{n\times n}_{sym}$ is of class $C^{1+\alpha}$. Let I_{n-1} denote the $n\times(n-1)$ unit matrix (i.e., the first $n-1$ columns of $\sum_{j=1}^{n-1} e_j \otimes e_j$)*

and let $E := I_{n-1} + e_n \otimes D\chi(x') - x_n D_{x'}\nu(x', \chi(x')) \in \mathbb{M}^{n\times(n-1)}$. *Then*

$$g(x) = \begin{pmatrix} E^T E & 0 \\ 0 & 1 \end{pmatrix} \in \mathbb{M}^{n\times n}_{sym} \text{ for } (x', x_n) \in B.$$

Proof. We have $D\Psi = E - \nu \otimes e_n$. Hence $\Psi \in C^{1+\alpha}$. Moreover $E^T E e_n = 0$ and

$$E^T \nu = \sum_{j=1}^{n} \big(\nu_j + (\partial_{x_j}\chi)\nu_n - x_n(\partial_{x_j}\nu)\cdot\nu_n\big)\, e_j.$$

Since ν is normal to the surface $x' \mapsto (x', \chi(x'))$ we have $(\nu_j + \partial_{x_j}\chi) = \partial_{x_j}(x', \chi(x'))\cdot \nu = 0$. Moreover $\partial_{x_j}\nu\cdot\nu = 0$ since $|\nu|^2 = 1$. This establishes the desired block structure of g. ∎

DEFINITION 4.2. *Suppose that* δ *is small enough,* $0 < \delta < \delta_0$, *such that* $\Psi(B_+) =: \omega \subset \Omega$, $B_\pm := \{(x', x_n) \in B : \pm x_n > 0\}$, *where* $B := B(0, \delta)$ *and* $B' := \{x' \in \mathbb{R}^{n-1} : |x'| < \delta\}$ *denote the* δ*-ball around* $x_0 = 0$ *in* n *and* $(n-1)$ *dimensions, respectively. For any* $x = (x', x_n) \in B_+$ *set* $Sx := (x', -x_n) \in B_-$ *and*

$$\begin{aligned}
\tilde u(x) &= -\tilde u(Sx) := u(\Psi(x)),\\
\tilde\sigma(x) &= -\tilde\sigma(Sx)S := \sigma(\Psi(x))\ \operatorname{cof} D\Psi,\\
\tilde f(x) &= -\tilde f(Sx) := (\det g(x))^{1/2}\, f(\Psi(x)),\\
\tilde g(x) &= \tilde g(Sx) := g(x).
\end{aligned}$$

LEMMA 4.5. *There holds* $\tilde u \in W^{1,2}(B)$, $\tilde f \in W^{1,2}(B)$, $\tilde\sigma \in H(\operatorname{div}, B)$,

$$\tilde\sigma = D\varphi(\nabla\tilde u\, \tilde g^{-1/2}) \operatorname{cof} \tilde g^{1/2} \text{ in } B, \quad \text{and} \quad \operatorname{div}\tilde\sigma = \tilde f \text{ in } \mathcal{D}'(B).$$

Proof. A polar decomposition $QU = D\Psi(x)$ shows that $g(x) = U^2$, $g(x)^{1/2} = U$. Since $Q = D\Psi(x) g(x)^{-1/2}$ is orthonormal we have

$$|\nabla u(\Psi(x))| = |\nabla u(\Psi(x))\, D\Psi(x)\, g(x)^{-1/2}| = |\nabla\tilde u(x)\, g(x)^{-1/2}|.$$

The function $\varphi(\cdot) = \psi(|\cdot|)$ solely depends on the modulus, and this implies that for $\xi := \Psi(x)$ and $x \in B_+$,

$$\begin{aligned}
\sigma(\xi) &= D\varphi(\nabla u(\xi))\\
&= \psi'(|\nabla u(\xi)|)\, \nabla u(\xi)/|\nabla u(\xi)|\\
&= \psi'(|\nabla\tilde u(x)\, g(x)^{-1/2}|)(\nabla\tilde u(x)\, D\Psi^{-1}(x))/|\nabla\tilde u(x)\, D\Psi^{-1}(x)|\\
&= D\varphi(\nabla\tilde u(x)\, g^{-1/2}(x))\, Q^T.
\end{aligned}$$

Using $\operatorname{adj} D\psi(x) = \operatorname{cof} g^{1/2}(x)\, Q^T$ we obtain the asserted identity for $\tilde\sigma$ in B_+.

By assumption $\operatorname{div}\sigma = f$ in $\mathcal{D}'(\Omega)$. Using the test function $\eta\circ\Psi^{-1}$, where $\eta \in \mathcal{D}(B_+)$, and the change of variables formula we get

$$\begin{aligned}
\int_{B_+} \tilde f(x)\, \eta(x)\, dx &= \int_\omega f(\xi)\eta(\Psi^{-1}(\xi))\, d\xi\\
&= -\int_\omega \nabla\eta(\Psi^{-1}(\xi))\cdot\big(D\Psi^{-1}(\xi)\sigma(\xi)\big)\, d\xi.
\end{aligned}$$

The substitution of $\tilde\sigma$ and a retransformation give

$$\int_{B_+} \tilde f(x)\, \eta(x)\, dx = \int_{B_+} \tilde\sigma(x)\cdot\nabla\eta(x)\, dx.$$

This proves $\operatorname{div}\tilde\sigma = \tilde f$ in $\mathcal{D}'(B_+)$.

The block structure of g shows that g^α commutes with $S = \operatorname{diag}(1,\dots,1,-1)$, i.e., $Sg^\alpha = g^\alpha S$ for $\alpha\in\mathbb{R}$. Since $\varphi(\cdot)$ depends solely on the modulus, $D\varphi$ commutes with S as well, i.e., $D\varphi(-S\cdot) = -SD\varphi(\cdot)$. Then, for $x\in B_-$, $\xi\in B_+$, $x = S\xi$,

$$\begin{aligned}&\operatorname{cof}\tilde g^{1/2}(x)\, D\varphi(\tilde g^{-1/2}(x)\,\nabla\tilde u(x))\\ &= \operatorname{cof}\tilde g^{1/2}(\xi)\, D\varphi(-\tilde g^{-1/2}(\xi)\, S\nabla\tilde u(\xi))\\ &= -S\operatorname{cof}\tilde g^{1/2}(\xi)\, D\varphi(\tilde g^{-1/2}(\xi)\,\nabla\tilde u(\xi))\\ &= -\tilde\sigma(\xi)S = \tilde\sigma(x).\end{aligned}$$

Thus $\tilde\sigma = \operatorname{cof}(\tilde g^{1/2})\, D\varphi(\tilde g^{-1/2}\,\nabla\tilde u)$ holds a.e. in B.

Since $\tilde f = 0 = \tilde u$ on $\overline{B}_+\cap\overline{B}_- = B\cap(B'\times\{0\})$ the maps $\tilde u$ and $\tilde f$ belong to $W^{1,2}(B)$. Notice that $g\in C(B)$. Clearly $\tilde\sigma\in L^2(B)$ and $\tilde\sigma|_{B_\pm}\in H(\operatorname{div},B_\pm)$. Hence it remains to prove $\operatorname{div}\tilde\sigma = \tilde f$ in $\mathcal{D}'(B)$. Given $\eta\in\mathcal{D}(B)$, set $\alpha := (\eta+\eta\circ S)/2$ and $\beta := (\eta-\eta\circ S)/2$. Since $\nabla\alpha(x) = (\nabla\eta(x)+\nabla\eta(Sx)S)/2 = \nabla\alpha(Sx)S$ we get

$$\int_B \tilde\sigma\cdot\nabla\alpha\,dx = \int_{B_+}(\tilde\sigma(x)+\tilde\sigma(Sx))\cdot\nabla\alpha(x)\,dx = 0.$$

We have $\beta = 0$ on $B'\times\{0\}$ and $\nabla\beta(Sx) = -\nabla\beta(x)S$. Thus a transformation to B_+ and an integration by parts in B_+ lead to

$$\begin{aligned}&\int_B \tilde\sigma\cdot\nabla\eta\,dx = \int_{B_+}\tilde\sigma(x)\cdot\nabla\beta(x)\,dx + \int_{B_+}\tilde\sigma(Sx)\cdot\nabla\beta(Sx)\,dx\\ &= 2\int_{B_+}\tilde\sigma(x)\cdot\nabla\beta(x)\,dx = 2\int_{B_+}\tilde f(x)\beta(x)\,dx = \int_B\tilde f(x)\beta(x)\,dx.\end{aligned}$$

Hence $-\operatorname{div}\tilde\sigma = \tilde f$ in $\mathcal{D}'(B)$, and the proof of Lemma 4.5 is finished. □

If we can show that the transformed stress $\tilde\sigma$ satisfies $\tilde\sigma\in W^{1,2}_{loc}(B;\mathbb{M}^{n\times n})$, then we easily conclude that $\sigma\in W^{1,2}(\Omega\cap B(x_0,\delta/2);\mathbb{R}^n)$, and we are done. To establish the regularity of $\tilde\sigma$ we cannot directly appeal to Theorem 2.1 since $\tilde\sigma$ is not of the form $D\varphi(\nabla\tilde u)$ but is given by the inhomogeneous expression $\operatorname{cof}(\tilde g^{1/2})\, D\varphi(\tilde g^{-1/2}\,\nabla\tilde u)$. To conclude we thus adopt the strategy of the proof of Theorem 2.1 for the present situation with variable coefficients.

Proof of Theorem 4.3. Let $x\in B$, $h>0$, and let $M\in\mathbb{R}^n$ with $|M| = 1$. We use the abbreviations $x_2 := x+hM$, $x_0 := x-h\,M$, $x_1 := x$ and define, for $j = 1,2$,

$$\begin{aligned}&F_j := \nabla\tilde u(x_j),\quad \sigma_j := \tilde\sigma(x_j),\quad U_j := \tilde g^{-1/2}(x_j),\\ &V_j := \operatorname{cof}U_j^{-1},\quad \Sigma_j := \sigma_j\,\det U_j,\quad T_j := \Sigma_jU_j^{-1}.\end{aligned}$$

We write $a\lesssim b$ to denote $a\le C\,b$ if C is a generic constant that is independent of $\delta>0$, $h>0$ (as long as they are sufficiently small). The constant $C>0$ may, however, depend on g, U_j, V_j, e.g., through $\|\tilde g\|_{W^{1,\infty}(B)}$, $\|\operatorname{cof}\tilde g\|\|_{W^{1,\infty}(B)}$, $\|\tilde g^{-1}\|_{W^{1,\infty}(B)}$, $\|\operatorname{cof}\tilde g^{-1}\|_{W^{1,\infty}(B)}$, or $\|\eta\|_{W^{1,\infty}(B)}$. With this notation we have

$$\begin{aligned}|\sigma_2-\sigma_1|^2 &= |D\varphi(F_2U_2)V_2 - D\varphi(F_1U_1)V_1)|^2\\ &\lesssim |V_2-V_1|^2|D\varphi(F_1U_1)|^2 + |D\varphi(F_2U_2)-D\varphi(F_1U_1)|^2.\end{aligned}$$

Using Proposition 4.1 and the identity $T_jU_j = \Sigma_j$ we infer

$$\begin{aligned}
&|D\varphi(F_2U_2) - D\varphi(F_1U_1)|^2 \\
&\lesssim (D\varphi(F_2U_2) - D\varphi(F_1U_1)) \cdot (F_2U_2 - F_1U_1) \\
&= (T_2 - T_1) \cdot (F_2U_2 - F_1U_1) \\
&= (\Sigma_2 - \Sigma_1) \cdot (F_2 - F_1) + (T_2 - T_1) \cdot (F_1 + F_2)(U_2 - U_1) \\
&\quad + T_1 \cdot F_1(U_2 - U_1) - T_2 \cdot F_2(U_2 - U_1).
\end{aligned}$$

Consider $\eta \in \mathcal{D}(B)$, $0 \le \eta \le 1$, which equals one in a neighborhood of $x_0 = 0$, and assume that $|h|$ is sufficiently small. Multiply the combination of the last two estimates by η^2/h^2 and integrate over $\operatorname{supp}\eta$. With the notation $\tilde\tau(x) := (\tilde\sigma(x_2) - \tilde\sigma(x_1))/h$ and $\tilde e(x) := (\tilde u(x_2) - \tilde u(x_1))/h$, we deduce that

$$\begin{aligned}
&\int \eta^2(x)|\tilde\tau(x)|^2\,dx \lesssim \int \eta^2|\tilde\sigma(x)|^2\,dx \\
&+ 1/h^2 \int \eta^2(\Sigma_2 - \Sigma_1) \cdot (F_2 - F_1)\,dx \\
&+ 1/h \int \eta^2|T_2 - T_1|\,(|\nabla\tilde u(x)| + |\nabla\tilde u(x + h\,M)|)\,dx \\
&+ 1/h^2 \int \eta^2 T_1(U_2 - U_1) \cdot F_1\,dx \\
&- 1/h^2 \int \eta^2 T_2(U_2 - U_1) \cdot F_2\,dx \\
&=: \text{ I } + \text{ II } + \text{ III } + \text{ IV} - \text{V}.
\end{aligned}$$

Term I is bounded since $\tilde\sigma \in L^2(B)$. Term II is recast into the form

$$\begin{aligned}
\text{II} &= 1/h^2 \int \eta^2(\Sigma_2 - \Sigma_1) \cdot (F_2 - F_1)\,dx \\
&= \int \eta^2(x)\,\det\tilde g^{-1/2}(x)\,\tilde\tau(x) \cdot \nabla\tilde e(x)\,dx \\
&\quad + \int \eta^2(x)\left(\det\tilde g^{-1/2}(x_2) - \det\tilde g^{-1/2}(x_1)\right)/h \\
&\qquad \times \tilde\sigma(x_2) \cdot \nabla\tilde e(x)\,dx.
\end{aligned}$$

Since $\tilde e\eta^2\,\det\tilde g^{-1/2} \in H^1(B)$ is a feasible test function we have

$$\begin{aligned}
&\int \eta^2 \det\tilde g^{-1/2}\tilde\tau \cdot \nabla\tilde e\,dx = -\int \tilde e\tilde\tau \cdot \nabla(\eta^2 \det\tilde g^{-1/2})\,dx \\
&+ \int \tilde e(x)\eta^2(x) \det g^{-1/2}(x)\,(f(x_2) - f(x_1))\,/h\,dx \\
&\lesssim \|\tilde u\|_{1,2}(\|f\|_{1,2} + \|\eta\tilde\tau\|_2)
\end{aligned}$$

with the abbreviations $\|\cdot\|_p := \|\cdot\|_{L^p(B)}$ and $\|\cdot\|_{1,p} := \|\cdot\|_{W^{1,p}(B)}$. To estimate the other term in II we write out $\nabla\tilde e(x) = 1/h\,(\nabla\tilde u(x + hM) - \nabla\tilde u(x))$, split the integral into two integrals, and perform a change of variables $x \mapsto x - hM$ in the first integral

(summation by parts). This yields

$$\begin{aligned}
&\int \eta^2(x)\big(\det \tilde g^{-1/2}(x_2) - \det \tilde g^{-1/2}(x_1)\big)/h\,\tilde\sigma(x_2)\cdot\nabla\tilde e(x)\,dx\\
&= -\int (\eta^2(x_1)-\eta^2(x_0))/h\\
&\qquad \times\ (\det \tilde g^{-1/2}(x_1) - \det \tilde g^{-1/2}(x_0))/h\,\tilde\sigma(x)\cdot\nabla\tilde u(x)\,dx\\
&\quad -\int \eta^2(x)(\det \tilde g^{-1/2}(x_2) - \det \tilde g^{-1/2}(x))/h\,\tilde\tau(x)\cdot\nabla\tilde u(x)\,dx\\
&\quad -\int \eta^2(x)\Big(\det \tilde g^{-1/2}(x_2) - 2\det \tilde g^{-1/2}(x_1)\\
&\qquad\qquad + \det \tilde g^{-1/2}(x_0)\Big)/h^2\,\tilde\sigma(x)\cdot\nabla\tilde u(x)\,dx\\
&\lesssim (\|\tilde\sigma\|_2 + \|\eta\tilde\tau\|_2)\cdot\|\tilde u\|_{1,2}.
\end{aligned}$$

In the last step we used $g\in C^{1,1}$. It is at this point that the assumption $\partial\Omega$ is $C^{2,1}$ enters. Combining the previous estimates, we get

$$\text{II} \ \lesssim \|\tilde u\|_{1,2}(\|f\|_{1,2} + \|\tilde\sigma\|_{1,2} + \|\eta\tilde\tau\|_2).$$

Since $T_j = D\varphi(\nabla\tilde u(x_j)) = \sigma_j \operatorname{cof} U_j$, similar arguments lead to

$$\begin{aligned}
\text{III} \ &= 1/h\int \eta^2|T_2 - T_1|(|\nabla\tilde u(x_1)| + |\nabla\tilde u(x_2)|)\,dx\\
&\lesssim (\|\eta\tilde\tau\|_2 + \|\tilde\sigma\|_2)\|u\|_{1,2}.
\end{aligned}$$

A shift in the variable x_2 in term V leads similarly to the estimate

$$\begin{aligned}
\text{IV} - \text{V} &= \int \eta^2 T_1(\tilde g^{-1/2}(x_2) - 2\tilde g^{-1/2}(x_1) + \tilde g^{-1/2}(x_0)/h^2\cdot\nabla\tilde u(x)\,dx\\
&\quad + \int (\eta^2(x_1) - \eta^2(x_2)/h\,T_1\,(\tilde g^{-1/2}(x) - \tilde g^{-1/2}(x_0)/h\cdot\nabla\tilde u(x)\,dx\\
&\lesssim \|\tilde\sigma\|_2\,\|\tilde u\|_{1,2}.
\end{aligned}$$

Absorbing $\|\eta\tau\|_2$ in terms II and III concludes the proof. ∎

5. A symmetric variant for geometrically linear models. This section concerns a variation of Theorem 2.1 for symmetrized gradients. Some (geometrically) linear models in elasticity involve the symmetric Green strain

$$\varepsilon(u) := \operatorname{sym} Du := ((\partial u_j/\partial x_k + \partial u_k/\partial x_j) :\ j,k = 1,\dots,n)$$

for $m = n$. For ease of presentation we focus on $p = r = q = 2$ and $s = 0$ as in linear elasticity but emphasize robustness with respect to the limit $\lambda\to\infty$, where λ is one of the Lamé constants (see below).

Let $\mathbb{M}^{n\times n}_{sym}$ denote the symmetric real $n\times n$ matrices. The fourth-order elasticity tensor $\mathbb{C}:\mathbb{M}^{n\times n}_{sym}\to\mathbb{M}^{n\times n}_{sym}$ is defined by

$$\mathbb{C}E := \lambda\ \operatorname{tr}(E)\,\mathbb{I} + 2\mu\,E \quad \text{for } E\in\mathbb{M}^{n\times n}_{sym}.$$

Here the positive scalars λ,μ are the Lamé constants, $\mathbb{I}$ is the identity matrix, and $\operatorname{tr}(E) := \sum_{j=1}^n E_{jj}$. Since $\mathbb{C}$ is symmetric and positive definite, there exist an inverse $\mathbb{C}^{-1}$ and the square roots $\mathbb{C}^{1/2}$ and $\mathbb{C}^{-1/2}$. The norm

$$|\,E\,|_{\mathbb{C}} := (E : \mathbb{C}E)^{1/2} = |\,\mathbb{C}^{1/2}E\,| \quad \text{for } E\in\mathbb{M}^{n\times n}_{sym}$$

is induced by the energy scalar product with respect to $\mathbb{C}$ in $\mathbb{M}^{n\times n}_{sym}$.

Suppose that $\varphi : \mathbb{M}^{n\times n}_{sym} \to \mathbb{R}$ is C^1 and that there exists a constant c_5 such that for all $A, B \in \mathbb{M}^{n\times n}_{sym}$,

$$|D\varphi(A) - D\varphi(B)|^2_{\mathbb{C}^{-1}} \le c_5\,(D\varphi(A) - D\varphi(B)) : (A - B). \tag{5.1}$$

THEOREM 5.1. *Assume furthermore that*

$$u \in H^1(\Omega;\mathbb{R}^n) \quad \textit{and} \quad \sigma := D\varphi(\varepsilon(u))$$

satisfy

$$\sigma \in L^2_{loc}(\Omega;\mathbb{M}^{n\times n}_{sym}) \quad \textit{and} \quad \operatorname{div}\sigma \in H^1_{loc}(\Omega;\mathbb{R}^n).$$

Then

$$\sigma \in H^1_{loc}(\Omega;\mathbb{M}^{n\times n}_{sym}).$$

Moreover, if $\omega_0 \subset\subset \omega_1 \subset\subset \Omega$ for nonvoid open sets ω_0 and ω_1, there exists a λ-independent constant $c_6 > 0$ such that

$$\|\,\sigma\,\|_{H^1(\omega_0)} \le c_6\,(\|\,u\,\|_{H^1(\omega_1)} + \|\,\sigma\,\|_{L^2(\omega_1)} + \|\,\operatorname{div}\sigma\,\|_{H^1(\omega_1)}). \tag{5.2}$$

Remark 5.1. (a) Korn's inequality does not play an explicit role in the proof. It is used, however, in applications to guarantee $u \in H^1(\Omega)$ (and so the boundedness of e in $L^2_{loc}(\Omega)$ in the proof).

(b) The fourth-order elasticity tensor could be more general; for the assertion $\sigma \in H^1_{loc}(\Omega;\mathbb{M}^{n\times n}_{sym})$ it is sufficient that $\mathbb{C}$ is a linear, continuous, and positive definite operator.

(c) The constant c_6 in (5.2) depends on c_5, μ, ω_0, and ω_1, but on neither σ nor u.

(d) The functional $\varphi : \mathbb{M}^{n\times n}_{sym} \to \mathbb{R}$ may depend on $\mathbb{C}$ and λ; the constant c_6 depends on φ only through c_5 and stays independent of λ as long as c_5 does.

Proof of Theorem 5.1. The proof follows the arguments of the proof of Theorem 2.1, but the differential operator D is replaced by the symmetric variant ε. In particular, $\delta := \varepsilon(e)$. This results in the estimate

$$\begin{aligned}
\|\,\eta\,\mathbb{C}^{-1/2}\tau\,\|_2^2 &\le c_7 \int_\omega \eta^2\,\varepsilon(e) : \tau\,dx = -c_7 \int_\omega e\,\operatorname{div}(\tau\eta^2)\,dx \\
&\le c_7\,\|\,e\,\|_2\,\|\,\eta\,\|_{W^{1,\infty}(\Omega)}(2\|\,\eta\,\tau\,\|_2 + \|\,\eta\,\operatorname{div}\tau\,\|_2) \\
&\le c_8(\|\,u\,\|^2_{H^1(\omega)} + \|\,\operatorname{div}\sigma\,\|^2_{H^1(\omega)} + \|\,\eta\,\tau\,\|_2)^2
\end{aligned} \tag{5.3}$$

for some (h,λ,μ)-independent constant $c_8 > 0$. The first assertion follows (with a constant depending on λ) from (5.3) and the estimate

$$\|\,\eta\,\tau\,\|_2 \le (2\mu + \lambda)^{1/2}\|\,\eta\,\mathbb{C}^{-1/2}\tau\,\|_2.$$

In order to prove (5.2) fix $\omega_0 \subset\subset \omega_1$ and suppose that ω is a bounded Lipschitz domain between ω_0 and ω_1, $\omega_0 \subset\subset \omega \subset\subset \omega_1$. Assume that $\eta \in \mathcal{D}(\omega)$ satisfies $0 \le \eta \le 1$ and equals $\eta = 1$ on ω_0. Then we introduce the deviator $\operatorname{dev}(\tau) := \tau - \operatorname{tr}(\tau)/n\,\mathbb{I}$ and rewrite (5.3) as

$$\begin{aligned}
\|\,\eta\,\mathbb{C}^{-1/2}\tau\,\|_2^2 &= \frac{\|\,\operatorname{dev}(\eta\tau)\,\|_2^2}{2\mu} + \frac{\|\,\operatorname{tr}(\eta\tau)\,\|_2^2}{n^2(2\mu/n + \lambda)} \\
&\le c_8\Big(\|\,u\,\|^2_{H^1(\omega)} + \|\,\operatorname{div}\sigma\,\|^2_{H^1(\omega)} + \|\,\operatorname{dev}(\eta\tau)\,\|_2^2 + \|\,\operatorname{tr}(\eta\tau)\,\|_2^2\Big).
\end{aligned} \tag{5.4}$$

The λ-independent bound requires an extra argument using results for the Stokes problem [BF91, GR86]. To apply this we first reduce to a situation with zero mean. Let ξ be the center of mass of ω. We write $|\omega|$ for the measure of ω, and we set $e_1 = (1, 0, \dots, 0)$. Define the constant

$$\tau_0 := \int_\omega \operatorname{tr}(\eta\,\tau) dx/|\omega| \in \mathbb{R}$$

and consider the function $v_1 \in H^1(\omega; \mathbb{R}^n)$ defined by

$$v_1(x) := \tau_0((x - \xi) \cdot e_1)\, e_1 \quad \text{for } x \in \omega.$$

Then $\operatorname{div} v_1 = \tau_0$ and $\int_\omega (\tau_0 - \operatorname{tr}(\eta\,\tau))dx = 0$. The solvability of the Stokes equations guarantees the existence of $v_2 \in H^1_0(\omega; \mathbb{R}^n)$, which satisfies $\operatorname{div} v_2 = \tau_0 - \operatorname{tr}(\eta\tau)$ and the bound

$$\| v_2 \|_{H^1(\omega)} \le c_9 \| \tau_0 - \operatorname{tr}(\eta\,\tau) \|_2 \le c_9 \| \operatorname{tr}(\eta\,\tau) \|_2.$$

Thus there exists $c_{10} > 0$ such that $v := v_1 - v_2 \in H^1(\omega; \mathbb{R}^n))$ satisfies

$$\operatorname{div} v = \operatorname{tr}(\eta\,\tau) \quad \text{and} \quad \| v \|_{H^1(\omega)} \le c_{10} \| \operatorname{tr}(\eta\,\tau) \|_2. \tag{5.5}$$

Recall that $\operatorname{tr}(\eta\,\tau)/n\,\mathbb{I} := \eta\,\tau - \operatorname{dev}(\eta\,\tau)$. Hence

$$\begin{aligned} \| \operatorname{tr}(\eta\,\tau) \|_2^2 &= \int_\omega \operatorname{tr}(\eta\,\tau) \operatorname{div} v\, dx = \int_\omega \operatorname{tr}(\eta\,\tau)\mathbb{I} : Dv\, dx \\ &= n \int_\omega (\eta\,\tau - \operatorname{dev}(\eta\,\tau)) : Dv\, dx. \end{aligned}$$

Multiple applications of Cauchy's inequality and integration by parts yield

$$\begin{aligned} \frac{1}{n} \| \operatorname{tr}(\eta\,\tau) \|_2^2 &\le \| Dv \|_2 \| \operatorname{dev}(\eta\,\tau) \|_2 - \int_\Omega v \cdot (\tau \nabla\eta + \eta \operatorname{div}\tau) dx \\ &\le \| v \|_{H^1(\omega)} \Big(\| \operatorname{dev}(\eta\,\tau) \|_2 + \| \eta \operatorname{div}\tau \|_2 \Big) \\ &\quad - \int_\Omega v \cdot \tau \nabla\eta\, dx. \end{aligned} \tag{5.6}$$

To rewrite the last term using a summation by parts, let $\otimes$ denote the dyadic product and set

$$V_h(x) := \frac{1}{h} \Big((v \otimes \nabla\eta)(x) - (v \otimes \nabla\eta)(x - hM) \Big) \in \mathbb{M}^{n\times n} \quad \text{for a.e. } x \in \omega.$$

Now $(v \otimes \nabla\eta)_{jk} = v_j\, \partial\eta/\partial x_k$ belongs to $H^1(\omega)$, and we have

$$\lim_{h\to 0} \| V_h \|_2 \le | v \otimes \nabla\eta |_{H^1(\omega_1)} \le \| v \|_{H^1(\omega)} \| \eta \|_{W^{2,\infty}(\omega)}. \tag{5.7}$$

Since $\eta \in \mathcal{D}(\omega)$ is fixed we infer (for sufficiently small h) from (5.7) that

$$\begin{aligned} &- \int_\Omega v \cdot \tau \nabla\eta\, dx = \int_\Omega V_h : \sigma\, dx \\ &\le \| \sigma \|_{L^2(\omega_1)} \| V_h \|_{L^2(\omega_1)} \le c_{11} \| \sigma \|_{L^2(\omega_1)} \| v \|_{H^1(\omega)}. \end{aligned}$$

Using this in (5.6) and applying the estimate (5.5) to bound $\| v \|_{H^1(\omega)}$, we deduce

$$c_{12} \| \operatorname{tr}(\eta\,\tau) \|_2 \le \| \operatorname{dev}(\eta\,\tau) \|_2 + \| \operatorname{div} \tau \|_2 + \| \sigma \|_{L^2(\omega_1)}. \tag{5.8}$$

We return to (5.4) and substitute $\| \operatorname{tr}(\eta\,\tau) \|_2$ with the bound (5.8) on the right-hand side of (5.4). The resulting estimate reads

$$\begin{aligned} &\frac{\| \operatorname{dev}(\eta\tau) \|_2^2}{2\mu} + \frac{\| \operatorname{tr}(\eta\tau) \|_2^2}{n^2(2\mu/n+\lambda)} \\ &\le c_{13}\Big(\| u \|_{H^1(\omega)}^2 + \| \operatorname{div} \sigma \|_{H^1(\omega)}^2 + \| \sigma \|_{L^2(\omega_1)}^2 + \| \operatorname{dev}(\eta\tau) \|_2^2 \Big) \end{aligned}$$

and allows us to absorb $\| \operatorname{dev}(\eta\tau) \|_2$ using Young's inequality. Hence

$$c_{14} \| \eta\, \mathbb{C}^{-1/2} \tau \|_2 \le \| u \|_{H^1(\omega)} + \| \operatorname{div} \sigma \|_{H^1(\omega)} + \| \sigma \|_{L^2(\omega_1)}.$$

Another application of (5.8) finally yields

$$c_{15} \| \eta\, \tau \|_2 \le \| u \|_{H^1(\omega)} + \| \operatorname{div} \sigma \|_{H^1(\omega)} + \| \sigma \|_{L^2(\omega_1)}.$$

The proof is then concluded as in Theorem 2.1. □

6. An application to Hencky elastoplasticity with hardening. One time step within an elastoplastic evolution problem leads to Hencky's model. For various hardening laws and von-Mises yield conditions, the minimization problem takes the form (1.4). After an elimination of internal variables [ACZ99] the energy function becomes, in the notation of the previous section,

$$\varphi(E) := \frac{1}{2}\, E : \mathbb{C}E - \frac{1}{4\mu}\, \max\{0, |\operatorname{dev} \mathbb{C}E| - \sigma_y\}^2/(1+\eta) \tag{6.1}$$

for $E \in \mathbb{M}^{n\times n}_{sym}$. Here $\mathbb{C}$ is the fourth-order elasticity tensor, $\sigma_y > 0$ is the yield stress, and $\eta > 0$ is the modulus of hardening. The model of perfect plasticity corresponds to $\eta = 0$ [Tem83].

PROPOSITION 6.1. *We have, for all $A, B \in \mathbb{M}^{n\times n}_{sym}$,*

$$|D\varphi(A) - D\varphi(B)|^2_{\mathbb{C}^{-1}} \le (D\varphi(A) - D\varphi(B)) : (A - B). \tag{6.2}$$

Proof. Set $\xi(x) := 1 - \max\{0, 1 - \sigma_y/(2\mu\, x)\}/(1+\eta)$ to define the continuous and monotonously decreasing function $\xi : [0,\infty) \to (0,1]$ with $\xi(0) = 1 \ge \xi(x) > \eta/(1+\eta) > 0$ for $0 < x < \infty$. Then

$$D\varphi(E) = (\lambda + 2\mu/n)\, \operatorname{tr}(E)\, \mathbb{I} + 2\mu\, \xi(|\operatorname{dev} E|)\, \operatorname{dev} E \quad \text{for all } E \in \mathbb{M}^{n\times n}_{sym}.$$

Without loss of generality, we suppose that $a := |\operatorname{dev} A| \le b := |\operatorname{dev} B|$ and abbreviate $\alpha := \xi(a)$ and $\beta := \xi(b)$. First, we calculate

$$2\mu\, \delta := |D\varphi(A) - D\varphi(B)|^2_{\mathbb{C}^{-1}} - (D\varphi(A) - D\varphi(B)) : (A - B).$$

Then we have to show that

$$\delta = |\operatorname{dev}(\xi(a)A - \xi(b)B)|^2 - \operatorname{dev}(\xi(a)A - \xi(b)B) : \operatorname{dev}(A - B)$$

is nonpositive. To see that $\delta \le 0$ observe that $0 \le (1-\alpha)\beta + \alpha(1-\beta)$. Expanding the squares and collecting terms, we infer in combination with Cauchy's inequality that

$$\begin{aligned}\delta &= (\xi(a)a - \xi(b)b)^2 - (\xi(a)a - \xi(b)b)(a-b)\\ &\quad + (\operatorname{dev}(A) : \operatorname{dev}(B) - ab)\Big((1-\alpha)\beta + \alpha(1-\beta)\Big)\\ &\le (\xi(a)a - \xi(b)b)^2 - (\xi(a)a - \xi(b)b)(a-b)\\ &= (\xi(a)a - \xi(b)b)\Big((\alpha-1)a - (\beta-1)b\Big).\end{aligned}$$

An elementary analysis shows that $x\xi(x) \ge 0$ is monotonously increasing in $0 \le x < \infty$, while $x(\xi(x)-1) \le 0$ is monotonously decreasing. As a consequence, $a \le b$ implies $\xi(a)a \le \xi(b)b$ and $(\xi(a)-1)a \ge (\xi(b)-1)b$. Taking this into account in the last estimate of δ, we conclude that $\delta \le 0$. ∎

We therefore have the following consequence of Theorem 5.1.

COROLLARY 6.2. *If u is a minimizer of (1.2) in $\mathcal{A} \subseteq H^1(\Omega)$ and $f \in H^1_{loc}(\Omega)$, then $\sigma := D\varphi(\varepsilon(u)) \in W^{1,2}_{loc}(\Omega;\mathbb{R}^n)$.* ∎

Remark 6.1. (a) The corollary is essentially due to Seregin [Ser93].

(b) The case $\eta = 0$ corresponds to perfect plasticity [Tem83] and is excluded from our analysis. Then u belongs only to $BD(\Omega)$, the space of bounded deformations.

7. An application to a vector two-well problem. Given two distinct wells E_1 and E_2 in $\mathbb{M}^{n\times n}_{sym}$ with minimal energies W_1^0 and W_2^0 in $\mathbb{R}$, we consider the quadratic elastic energies

$$W_j(E) := \frac{1}{2}(E-E_j) : \mathbb{C}(E-E_j) + W_j^0 \quad \text{for all } E \in \mathbb{M}^{n\times n}_{sym}. \tag{7.1}$$

Energy minimization leads to an optimal choice of the configuration of the two phases, and so the strain energy density W is modeled by the minimum

$$W(E) = \min\{W_1(E), W_2(E)\} \quad \text{for all } E \in \mathbb{M}^{n\times n}_{sym}. \tag{7.2}$$

The two wells (transformation strains) are said to be compatible if the following condition holds:

$$E_1 = E_2 + \frac{1}{2}(a\otimes b + b\otimes a) \quad \text{for some } a, b \in \mathbb{R}^n. \tag{7.3}$$

The constant γ used below is determined by a certain projection onto the space of symmetric matrices and satisfies $0 < \gamma \le \frac{1}{2}|E_2 - E_1|^2_{\mathbb{C}}$. In the compatible case (7.3) it attains its upper bound, i.e., $\gamma = \frac{1}{2}|E_2 - E_1|^2_{\mathbb{C}}$. The quasi-convexification φ of W is given by [Koh91]:

$$\varphi(E) = \begin{cases} W_2(E) & \text{if } W_2(E) + \gamma \le W_1(E),\\ \frac{1}{2}(W_2(E)+W_1(E)) - \frac{1}{4\gamma}(W_2(E)-W_1(E))^2 - \frac{\gamma}{4} & \\ \qquad\qquad \text{if } |W_2(E) - W_1(E)| \le \gamma,\\ W_1(E) & \text{if } W_1(E) + \gamma \le W_2(E).\end{cases} \tag{7.4}$$

LEMMA 7.1 (see [CP97a]). *In the compatible case (7.3) we have, for all $A, B \in \mathbb{M}^{n\times n}_{sym}$,*

$$|D\varphi(A) - D\varphi(B)|^2_{\mathbb{C}^{-1}} \le \Big(D\varphi(A) - D\varphi(B)\Big) : (A-B). \tag{7.5}$$

Thus Theorem 5.1 implies the following result.

COROLLARY 7.2 (see [Ser96]). *If u is a minimizer of (1.4) in $\mathcal{A} \subseteq H^1(\Omega)$ and $f \in H^1_{loc}(\Omega)$, then $\sigma := D\varphi(\varepsilon(u)) \in W^{1,2}_{loc}(\Omega; \mathbb{M}^{n\times n})$.* □

Remark 7.1. (a) The corollary is due to Seregin [Ser96, Theorem 2.2]. In addition to the local stress regularity, he shows that the strain tensor locally has bounded mean oscillation, and he investigates the pure phase area.

(b) In the case of incompatible wells (i.e., if (7.3) fails) Lemma 7.1 fails (as it guarantees convexity of φ). Seregin [Ser99] showed that the quasi-convex envelope $\varphi(\operatorname{sym} F)$ can be rewritten as the sum of a convex function (which then satisfies an estimate of the form (7.5)) and a linear combination of second-order minors of F. The integral $\int_\Omega \operatorname{cof} Du\, dx$ depends only on the boundary values of u and can hence be neglected in the pure Dirichlet problem. Then, up to cofactor matrices of the gradient F, the stress belongs to $W^{1,2}_{loc}(\Omega; \mathbb{M}^{n\times n}_{sym})$. Formally one may interpret the term $\int_\Omega \operatorname{cof} Du\, dx$ as a constant pressure (as the model is in material coordinates). Such an interpretation is, however, doubtful if one keeps in mind that (7.1) is based on a linearization. Thus material and spatial coordinates coincide and incompressibility reads $\operatorname{div} u = 0$ and *not* $\det Du = 1$.

(c) A time-discretized model for hysteresis in [MTL] leads to a similar variational problem. From a stress estimate in [CP00] we obtain an analogue of Lemma 7.1 and can conclude $\sigma \in W^{1,2}_{loc}(\Omega; \mathbb{M}^{n\times n})$ as well.

REFERENCES

[ACZ99] J. ALBERTY, C. CARSTENSEN, AND D. ZARRABI, *Adaptive numerical analysis in primal elastoplasticity with hardening*, Comput. Methods Appl. Mech. Engrg., 171 (1999), pp. 175–204.

[BF91] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[BKK00] J.M. BALL, B. KIRCHHEIM, AND J. KRISTENSEN, *Regularity of quasiconvex envelopes*, Calc. Var. Partial Differential Equations, 11 (2000), pp. 333–359.

[Bol06] O. BOLZA, *A fifth necessary condition for a strong extremum of the integral* $\int_{x_0}^{x_1} f(x, y, y')dx$, Trans. Amer. Math. Soc., 7 (1906), pp. 314–324.

[Cel93a] A. CELLINA, *On minima of a functional of the gradient: Necessary conditions*, Nonlinear Anal., 20 (1993), pp. 337–341.

[Cel93b] A. CELLINA, *On minima of a functional of the gradient: Sufficient conditions*, Nonlinear Anal., 20 (1993), pp. 343–347.

[CK01] C. CARSTENSEN AND R. KLOSE, *Guaranteed a Posteriori Finite Element Error Control for the p-Laplace Problem*, preprint, 2001.

[CP97a] C. CARSTENSEN AND P. PLECHÁČ, *Adaptive algorithms for scalar non-convex variational problems*, Appl. Numer. Math., 26 (1997), pp. 203–216.

[CP97b] C. CARSTENSEN AND P. PLECHÁČ, *Numerical solution of the scalar double-well problem allowing microstructure*, Math. Comp., 66 (1997), pp. 997–1026.

[CP00] C. CARSTENSEN AND P. PLECHÁČ, *Numerical analysis of compatible phase transitions in elastic solids*, SIAM J. Numer. Anal., 37 (2000), pp. 2061–2081.

[Dac89] B. DACAROGNA, *Direct Methods in the Calculus of Variations*, Appl. Math. Sci. 78, Springer-Verlag, Berlin, 1989.

[Fri94] G. FRIESECKE, *A necessary and sufficient condition for non-attainment and formation of microstructure almost everywhere in scalar variational problems*, Proc. Roy. Soc. Edinburgh Ser. A, 124 (1994), pp. 437–471.

[GKR86] J. GOODMAN, R.V. KOHN, AND L. REYNA, *Numerical study of a relaxed variational problem from optimal design*, Comput. Methods Appl. Mech. Engrg., 57 (1986), pp. 107–127.

[GR86] V. GIRAULT AND P.A. RAVIART, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[Koh91] R.V. KOHN, *The relaxation of a double-well energy*, Contin. Mech. Thermodyn., 3 (1991), pp. 193–236.

[KSW91] B. Kawohl, J. Stara, and G. Wittum, *Analysis and numerical studies of a problem of shape design*, Arch. Ration. Mech. Anal., 114 (1991), pp. 349–363.

[MTL] A. Mielke, F. Theil, and V. I. Levitas, *A variational formulation of rate-independent phase transformations using an extremum principle*, Arch. Ration. Mech. Anal, 162 (2002), pp. 133–177.

[Rou97] T. Roubíček, *Relaxation in Optimization Theory and Variational Calculus*, de Gruyter Ser. Nonlinear Anal. Appl. 4, Walter de Gruyter, Berlin, 1997.

[Ser93] G.A. Seregin, *On the regularity of minimizers of some variational problems in the theory of plasticity*, St. Petersburg Math. J., 4 (1993), pp. 989–1020.

[Ser96] G.A. Seregin, *On the regularity properties of solutions of variational problems in the theory of phase transitions in an elastic body*, St. Petersburg Math. J., 7 (1996), pp. 979–1003.

[Ser99] G.A. Seregin, *A variational problem on the phase equilibrium of an elastic body*, St. Petersburg Math. J., 10 (1999), pp. 477–506.

[Tem83] R. Temam, *Problemes mathematiques en plasticite [Mathematical Problems in Plasticity]*, Mèthodes Mathematiques de l'Informatique [Mathematical Methods of Information Science] 12, Gauthier-Villars, Paris, 1983 (in French).

[You69] L.C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, London, Toronto, 1969.

SIAM J. MATH. ANAL.
Vol. 34, No. 3, pp. 511–526

ON A NONLINEAR PARTIAL DIFFERENTIAL EQUATION ARISING IN MAGNETIC RESONANCE ELECTRICAL IMPEDANCE TOMOGRAPHY*

SUNGWHAN KIM†‡, OHIN KWON§, JIN KEUN SEO†, AND JEONG-ROCK YOON¶

Abstract. This paper considers the fundamental questions, such as existence and uniqueness, of a mathematical model arising in the MREIT system, which is an electrical impedance tomography technique integrated with magnetic resonance imaging. The mathematical model for MREIT is the Neumann problem of a nonlinear elliptic partial differential equation $\nabla \cdot (\frac{a(x)}{|\nabla u(x)|}\nabla u(x)) = 0$. We show that this Neumann problem belongs to one of two cases: either infinitely many solutions exist or no solution exists. This explains rigorously the reason why we have used the modified model in [O. Kwon, E. J. Woo, J. R. Yoon, and J. K. Seo, *IEEE Trans. Biomed. Engrg.*, 49 (2002), pp. 160–167], which is a system of the Neumann problem associated with two different Neumann data. For this modified system, we prove a uniqueness result on the edge detection of a piecewise continuous conductivity distribution.

Key words. conductivity reconstruction, interior measurement, uniqueness, current density imaging, electrical impedance tomography, magnetic resonance imaging

AMS subject classifications. 35R30, 35J60, 31A25, 62P10, 92C55

PII. S0036141001391354

1. Introduction.

Magnetic resonance electrical impedance tomography (MREIT) is a new imaging technique of reconstructing the cross-sectional conductivity distribution of a human body by means of the electrical impedance tomography (EIT) technique integrated with the magnetic resonance imaging (MRI) technique. The EIT technique to estimate the conductivity distribution uses data obtained by injecting a known current into the body through electrodes placed on the surface and measuring the resulting voltage difference recorded on the electrodes. The EIT problem is known as a highly ill-posed inverse problem due to its low sensitivity of data to the change in conductivity value. (See [14].) MREIT is designed to overcome this severe ill-posedness of the EIT problem by making good use of a recent MRI technique, so-called current density imaging (CDI), which measures the internal current density distribution. (For related works see [4, 6, 10, 11, 12, 15].)

In the recent paper [7], a new reconstruction algorithm for MREIT was developed to provide a high-resolution conductivity image. This algorithm is based on a new mathematical modeling which is involved with a nonlinear partial differential equation instead of the linear conductivity equation. Although the algorithm has achieved

*Received by the editors June 25, 2001; accepted for publication (in revised form) May 8, 2002; published electronically December 13, 2002.

http://www.siam.org/journals/sima/34-3/39135.html

†Department of Mathematics, Yonsei University, Seoul 120–749, Korea. The first author's work was supported by the Brain Korea 21. The third author's work was supported by KOSEF grant 2001-2-103-001-5.

‡Current address: Department of Mathematical Sciences, University of Tokyo, 3-8-1 Komaba, Meguro, Tokyo 153, Japan (sungwhan@ms.u-tokyo.ac.jp).

§Department of Mathematics, Konkuk University, Seoul 143–701, Korea (oikwon@kkucc.konkuk.ac.kr). This author's work was supported by the Faculty Research Fund of Konkuk University in 2001.

¶School of Mathematics, Korea Institute for Advanced Study, Seoul 130-012, Korea. Current address: Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180 (yoonj@rpi.edu).

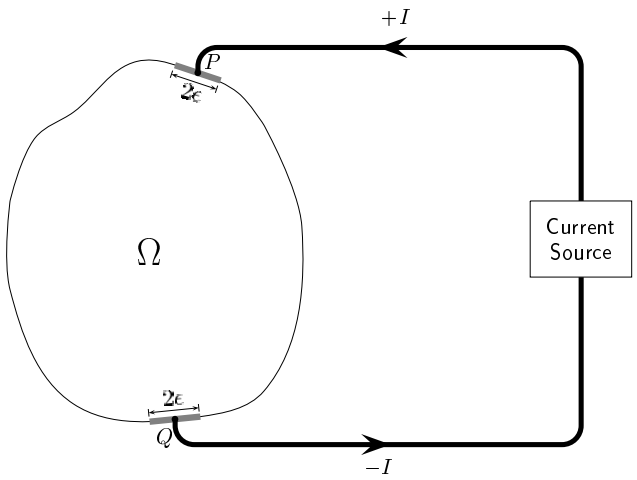


FIG. 1.1. *An illustration of the model where the current I is applied through a pair of electrodes attached to the boundary.*

successful numerical results in simulations, there has been no related mathematical theory for the new model such as existence and uniqueness. This paper is intended to provide answers to those questions.

Let us explain the mathematical model for MREIT, which was introduced in [7]. Let the cross-section of the cylindrical body occupy a bounded domain $\Omega \subset \mathbb{R}^2$. When a current is injected transversely through the outer surface of the body, it induces an electrical potential distribution u that satisfies the two-dimensional conductivity equation

$$\nabla \cdot (\sigma \nabla u) = 0 \quad \text{in } \Omega, \tag{1.1}$$

where σ denotes the conductivity coefficient of the body which we want to reconstruct. This unknown two variable function σ may be regarded as a piecewise continuous positive function. In the MREIT model, the current is applied through a pair of electrodes attached on the boundary $\partial\Omega$: If both electrodes of width 2ϵ are attached at points $P, Q \in \partial\Omega$, respectively, then the current density on the boundary can be approximated by a function

$$g(x) = \begin{cases} +\frac{I}{2\epsilon} & \text{on } \{|x-P| < \epsilon\} \cap \partial\Omega, \\ -\frac{I}{2\epsilon} & \text{on } \{|x-Q| < \epsilon\} \cap \partial\Omega, \\ 0 & \text{otherwise,} \end{cases} \tag{1.2}$$

where I is the current sent to both electrodes at P and Q; see Figure 1.1. For more details, see the ave-gab model in [3, 9].

With this current g, the resulting internal current density vector $\mathbf{J} = -\sigma\nabla u$ is divergence-free and satisfies the boundary condition

$$\sigma \frac{\partial u}{\partial \nu} = -\mathbf{J} \cdot \nu = g \qquad \text{on } \partial\Omega, \tag{1.3}$$

where ν denotes the outward unit normal vector to $\partial\Omega$. Moreover, the MREIT system furnishes the internal data $a = |\mathbf{J}| = \sigma|\nabla u|$, which is measured and processed in the

MRI system [7, 15]. We want to utilize this acquisition data a by substituting

$$\sigma(x) = \frac{a(x)}{|\nabla u(x)|}, \qquad x \in \Omega, \tag{1.4}$$

into the conductivity equation (1.1) and the Neumann boundary condition (1.3). As a result, the linear boundary value problem (1.1) and (1.3) with two unknowns σ and u is reduced to the following nonlinear Neumann boundary value problem with one unknown u:

$$\begin{aligned} \nabla\cdot\left(\frac{a}{|\nabla u|}\nabla u\right) &= 0 \quad \text{in } \Omega, \\ \frac{a}{|\nabla u|}\frac{\partial u}{\partial \nu} &= g \quad \text{on } \partial\Omega, \quad \text{and} \quad \int_{\partial\Omega} u\, ds = 0, \end{aligned} \tag{1.5}$$

where the last condition means the potential reference condition. To be precise, the electric potential $u \in H^1(\Omega)$ can be viewed as a weak solution satisfying

$$\int_\Omega \frac{a}{|\nabla u|}\nabla u \cdot \nabla\phi\, dx = \int_{\partial\Omega} g\phi\, ds \qquad \text{for all} \quad \phi \in H^1(\Omega)$$

with a constraint $\int_{\partial\Omega} u\, ds = 0$.

It is natural to investigate the fundamental mathematical issue of the nonlinear boundary value problem (1.5), such as existence and uniqueness. In practice, the existence may not be a serious problem, but the uniqueness must be seriously taken into account. In the case in which we have not one but several different solutions, there will be several corresponding distinct conductivity images, and we cannot judge which one would be the actual image.

Unfortunately, in section 3 we will prove that once (1.5) has a solution, then it always has infinitely many solutions under a practically acceptable assumption that will be precisely defined in section 3. Hence the model (1.5) using one measurement is insufficient for the reconstruction of the conductivity distribution. A numerical example is presented in section 5 to show how different conductivity images can be reconstructed with the same data (a, g). Moreover, we also prove in section 3 that (1.5) in general does not have an existence result even if a is smooth. We think that the existence of the solution to (1.5) is related to some complicated connection between a and g, because the internal current density a depends on the choice of the injected current g.

Thus the model should be modified in order to guarantee the uniqueness. In section 4, we apply two different currents g_1 and g_2 approximated in the same manner as in (1.2), attaching two different pairs of electrodes $\{P_1, Q_1\}$ and $\{P_2, Q_2\}$. Since the conductivity distribution σ is independent of the change of injected currents, from the relation (1.4) we may assume

$$\frac{a_1(x)}{|\nabla u_1(x)|} = \frac{a_2(x)}{|\nabla u_2(x)|}, \qquad x \in \Omega,$$

where u_j is a solution to the nonlinear Neumann boundary value problem (1.5) when g and a are replaced by g_j and a_j $(j = 1, 2)$. This leads to the following nonstandard

system of equations:

$$
\begin{aligned}
\nabla\cdot\left(\frac{a_j}{|\nabla u_j|}\nabla u_j\right) &= 0 \quad \text{in } \Omega,\\
\frac{a_1}{|\nabla u_1|} &= \frac{a_2}{|\nabla u_2|} \quad \text{in } \Omega,\\
\frac{a_j}{|\nabla u_j|}\frac{\partial u_j}{\partial \nu} &= g_j \quad \text{on } \partial\Omega,\\
\int_{\partial\Omega} u_j\, ds &= 0
\end{aligned}
\tag{1.6}
$$

for $j = 1, 2$. With this modified model and a practically acceptable assumption, in section 4 we are able to establish an important uniqueness result which may look strange at a glance.

In section 2, we define a space for physically meaningful conductivity distributions and recall some regularity properties of elliptic partial differential equations for further usage.

2. Definitions and preliminary. We assume that $\Omega \subset \mathbb{R}^2$, a cross-section of the human body, is a simply connected bounded domain with a $\mathcal{C}^2$ boundary. The conductivity distribution σ on the cross-section Ω may be regarded as a piecewise continuous function because distinct tissues have different conductivities. So, we may assume that σ belongs to the following class:

$$
\Sigma := \left\{ \sigma = \sigma_0 + \sum_{k=1}^{M} \sigma_k \chi_{D_k} \;\middle|\; M \in \mathbb{N},\ 0 < \sigma < \infty,\ \bar{D}_k \subset \Omega,\ \bar{D}_k \cap \bar{D}_\ell = \emptyset \text{ for } k \neq \ell, \right.
$$
$$
\left. \sigma_0 \in \mathcal{C}^\alpha(\bar{\Omega}),\ \sigma_k \in \mathcal{C}^\alpha(\bar{D}_k),\ \sigma_k \neq 0 \text{ on } \partial D_k,\ \partial D_k \text{ is a } \mathcal{C}^2 \text{ boundary} \right\},
$$

where χ_{D_k} denotes the characteristic function for D_k and $0 < \alpha < 1$ is not an important number. With this setting, for any $\sigma = \sigma_0 + \sum_{k=1}^{M} \sigma_k \chi_{D_k} \in \Sigma$, we easily see that

$$
\sigma \in \mathcal{C}^\alpha\left(\cup_{k=1}^{M} \bar{D}_k\right) \cap \mathcal{C}^\alpha\left(\Omega \setminus \cup_{k=1}^{M} D_k\right), \tag{2.1}
$$

$$
\{x \in \Omega \,|\, \sigma \text{ is discontinuous at } x\} = \bigcup_{k=1}^{M} \partial D_k. \tag{2.2}
$$

For a given current g in (1.2) and $\sigma = \sigma_0 + \sum_{k=1}^{M} \sigma_k \chi_{D_k} \in \Sigma$, let u be the solution of the classical Neumann boundary value problem

$$
\begin{aligned}
\nabla\cdot(\sigma\nabla u) &= 0 \quad \text{in } \Omega,\\
\sigma\frac{\partial u}{\partial \nu} &= g \quad \text{on } \partial\Omega, \quad \text{and} \quad \int_{\partial\Omega} u\, ds = 0.
\end{aligned}
\tag{2.3}
$$

From the basic theory of standard elliptic partial differential equations [5, 8], we know

$$
\begin{aligned}
&\text{(a)} \quad u \in \mathcal{C}(\bar{\Omega}),\\
&\text{(b)} \quad \nabla u \in \mathcal{C}^\alpha\left(\cup_{k=1}^{M} \bar{D}_k\right) \cap \mathcal{C}^\alpha\left(\Omega \setminus \cup_{k=1}^{M} D_k\right),\\
&\text{(c)} \quad \sigma_0(\xi)\nabla u^+(\xi)\cdot\nu(\xi) = (\sigma_0(\xi) + \sigma_k(\xi))\nabla u^-(\xi)\cdot\nu(\xi) \ \text{ if } \xi \in \partial D_k,\\
&\text{(d)} \quad \nabla u^+(\xi)\cdot\tau(\xi) = \nabla u^-(\xi)\cdot\tau(\xi) \ \text{ if } \xi \in \partial D_k,
\end{aligned}
\tag{2.4}
$$

where ν and τ are the outward unit normal vector and the unit tangent vector to ∂D_k, respectively, and u^+, u^- are defined by

$$u^+ = u|_{\Omega \setminus \bigcup_{k=1}^M \bar{D}_k} \quad \text{and} \quad u^- = u|_{\bigcup_{k=1}^M D_k}.$$

Moreover, owing to the choice of g as in (1.2), we can show that

$$\nabla u(x) \neq 0 \quad \text{for all } x \in \Omega, \tag{2.5}$$

the proof of which can be found in [1, 2, 13]. Indeed, (2.5) holds if nonzero g satisfies the following condition: there exist two disjoint arcs Γ^+ and Γ^- contained in $\partial\Omega$ such that

$$\Gamma^+ \cup \Gamma^- = \partial\Omega, \quad \text{and} \quad \Gamma^+ \subset \{g \geq 0\}, \quad \Gamma^- \subset \{g \leq 0\},$$

the detailed proof of which will be given in Remark 4.2 for completeness.

3. Nonexistence and nonuniqueness. In this section, we will prove that the nonlinear Neumann boundary value problem (1.5) under a practically acceptable assumption is generally not uniquely solvable by constructing infinitely many different solutions from one solution and by giving an example for nonexistence.

From the relation (1.4) between the conductivity distribution σ and the measured current density a, we may assume that a practically meaningful solution u of the Neumann problem (1.5) satisfies

$$\frac{a(x)}{|\nabla u(x)|} \in \Sigma, \tag{3.1}$$

since Σ contains almost all cases of piecewise continuous conductivities that may happen in the real situation. So, the practical solution u can be considered as a $H^1(\Omega)$ solution of the more complicated problem where g is given as in (1.2),

$$\begin{aligned} &\nabla\cdot\left(\frac{a}{|\nabla u|}\nabla u\right) = 0 \quad \text{in } \Omega, \quad \frac{a}{|\nabla u|} \in \Sigma, \\ &\frac{a}{|\nabla u|}\frac{\partial u}{\partial \nu} = g \quad \text{on } \partial\Omega, \quad \text{and} \quad \int_{\partial\Omega} u\, ds = 0. \end{aligned} \tag{3.2}$$

Hence, if u is a solution of (3.2), it satisfies (2.5) and the properties (a)–(d) in (2.4). By the property (b) in (2.4) and (3.1), $a = \frac{a}{|\nabla u|}|\nabla u|$ must be also a piecewise continuous function in Ω.

We can easily construct a solution for (3.2): For any $\sigma \in \Sigma$, there exists a unique solution u_σ to the classical Neumann problem (2.3), and this u_σ is also a solution to (3.2) when a is given by $a = \sigma|\nabla u_\sigma|$. To our surprise, (3.2) with this a has infinitely many solutions, and u_σ is just one of them. The following theorem states this nonuniqueness result.

THEOREM 3.1. *If the nonlinear problem* (3.2) *has a solution, then it has infinitely many solutions.*

Proof. Suppose u is a solution of the problem (3.2). We will construct infinitely many solutions by means of u. Since u satisfies the property (a) in (2.4) and (2.5), we have $\min_{x\in\bar{\Omega}} u(x) < \max_{x\in\bar{\Omega}} u(x)$. For any $t \in (\min_{\bar{\Omega}} u, \max_{\bar{\Omega}} u)$ and $\lambda > 0$, we define

$$u_{t,\lambda} := \begin{cases} u + c & \text{in} \quad \Omega_t^+, \\ \lambda u + (1-\lambda)t + c & \text{in} \quad \Omega_t^-, \end{cases}$$

where the number c is chosen so that $\int_{\partial\Omega} u_{t,\lambda}\, ds = 0$ and

$$\Omega_t^+ := \{x \in \Omega \,|\, u(x) \geq t\} \quad \text{and} \quad \Omega_t^- := \{x \in \Omega \,|\, u(x) < t\}.$$

Then it is easy to see that $u_{t,\lambda} \in \mathcal{C}(\bar{\Omega})$ and

$$\frac{\nabla u_{t,\lambda}(x)}{|\nabla u_{t,\lambda}(x)|} = \frac{\nabla u(x)}{|\nabla u(x)|} \quad \text{for all } x \in \Omega.$$

Since the possible discontinuity regions of $a/|\nabla u_{t,\lambda}|$ are $\{x \in \Omega \,|\, u(x) = t\}$ and those of $a/|\nabla u|$, we easily verify that $a/|\nabla u_{t,\lambda}| \in \Sigma$. Therefore $u_{t,\lambda}$ is also a solution to (3.2), which completes the proof. □

In section 5, we will present two distinct (numerically obtained) solutions that will arise in a complicated real situation and which solve the same problem (3.2).

Now we investigate the existence question. For simplicity, we confine ourselves to a unit square domain $\Omega = (0,1) \times (0,1)$ in $\mathbb{R}^2$. Let $x = (x_1, x_2)$ denote a point in Ω and let the current pattern g on $\partial\Omega$ be given by

$$g(x) = \begin{cases} -1 & \text{if } x_1 = 0, \\ 1 & \text{if } x_1 = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

The next theorem furnishes an example for the nonexistence of (3.2).

THEOREM 3.2. *Let $\Omega = (0,1) \times (0,1)$ and g be given in* (3.3)*. Assume that a in* (3.2) *depends only on the x_1-variable, that is, $a(x_1, x_2) = a(x_1)$. The necessary and sufficient condition for the existence of solution to* (3.2) *is $a \equiv 1$.*

Proof. If $a \equiv 1$, then clearly $u(x) = x_1 - \frac{1}{2}$ is a solution of (3.2) which would be one of the infinitely many solutions. This proves the sufficiency.

To show the necessity, suppose that the problem (3.2) has a solution u. First, we will prove that $a(t) \geq 1$ for all $t \in (0,1)$. For convenience, we denote

$$l_1 = \{x \in \partial\Omega \,|\, x_2 = 0\}, \qquad l_2 = \{x \in \partial\Omega \,|\, x_1 = 1\},$$
$$l_3 = \{x \in \partial\Omega \,|\, x_2 = 1\}, \qquad l_4 = \{x \in \partial\Omega \,|\, x_1 = 0\},$$

and let $R_t := \{x \in \Omega \,|\, 0 < x_1 < t\}$ be a rectangle on the left side of the line $\{x_1 = t\}$.

Applying the divergence theorem on R_t, we obtain

$$\begin{aligned} 0 &= \int_{\partial R_t} \frac{a}{|\nabla u|} \frac{\partial u}{\partial \nu}\, ds \\ &= \int_{\partial R_t \cap \partial\Omega} g\, ds + \int_{\partial R_t \cap \{x_1 = t\}} \frac{a}{|\nabla u|} \nabla u \cdot \nu\, ds \\ &= -\int_{l_4} ds + a(t) \int_{\partial R_t \cap \{x_1 = t\}} \frac{\nabla u \cdot \nu}{|\nabla u|}\, ds \\ &\leq -1 + a(t), \end{aligned} \tag{3.4}$$

since $a(x) = a(t)$ on $\partial R_t \cap \{x_1 = t\}$ and $\nabla u \cdot \nu \leq |\nabla u|$, where ν denotes the outward unit normal to R_t. Hence we have $a(t) \geq 1$ for all $t \in (0,1)$.

Now, we will show that the level curve $\Gamma_t := \{x \in \Omega \,|\, u(x) = u(t,0)\}$ is the vertical line $\{x \in \Omega \,|\, x_1 = t\}$ for all $t \in (0,1)$. Since $a \geq 1$, the choice of g in (3.3) and the Neumann boundary condition in (3.2) yield $\partial u/\partial\nu(t,0) = 0$, which implies

$$\min_{x \in \bar{\Omega}} u(x) < u(t,0) < \max_{x \in \bar{\Omega}} u(x).$$

Thus $\Omega_t := \{x \in \Omega \,|\, u(x) < u(t,0)\}$ is a nonempty open subset of Ω and $(t,0) \in \partial\Omega_t$. It is easy to see that

$$\mathcal{H}^1(\partial\Omega_t \cap l_4) \,<\, \mathcal{H}^1(\partial\Omega_t \setminus (l_1 \cup l_3 \cup l_4)) \quad \text{if } \Gamma_t \neq \{x \in \Omega \,|\, x_1 = t\}, \tag{3.5}$$

where $\mathcal{H}^1(L)$ denotes the arclength of the curve L.

Applying the divergence theorem on Ω_t, we have

$$\begin{aligned} 0 &= \int_{\partial\Omega_t} \frac{a}{|\nabla u|}\frac{\partial u}{\partial \nu}\, ds \\ &= \int_{\partial\Omega_t \cap \partial\Omega} g\, ds + \int_{\partial\Omega_t \setminus \partial\Omega} \frac{a}{|\nabla u|}\frac{\partial u}{\partial \nu}\, ds \\ &= -\mathcal{H}^1(\partial\Omega_t \cap l_4) + \mathcal{H}^1(\partial\Omega_t \cap l_2) + \int_{\partial\Omega_t \setminus \partial\Omega} a\frac{\nabla u \cdot \nu}{|\nabla u|}\, ds, \end{aligned} \tag{3.6}$$

where ν denotes the outward unit normal to Ω_t. Since $u(x) < u(t,0)$ in Ω_t and $u(x) = u(t,0)$ on $\partial\Omega_t \setminus \partial\Omega$, we have $\nu = \nabla u/|\nabla u|$ on $\partial\Omega_t \setminus \partial\Omega$, which implies

$$\frac{\nabla u \cdot \nu}{|\nabla u|} = 1 \quad \text{on } \partial\Omega_t \setminus \partial\Omega.$$

By the above identity and the fact that $a \geq 1$, from (3.6) we get

$$\begin{aligned} \mathcal{H}^1(\partial\Omega_t \cap l_4) &= \mathcal{H}^1(\partial\Omega_t \cap l_2) + \int_{\partial\Omega_t \setminus \partial\Omega} a\, ds \\ &\geq \mathcal{H}^1(\partial\Omega_t \cap l_2) + \mathcal{H}^1(\partial\Omega_t \setminus \partial\Omega) \\ &= \mathcal{H}^1(\partial\Omega_t \setminus (l_1 \cup l_3 \cup l_4)). \end{aligned}$$

Hence, from (3.5) it must be $\Gamma_t = \{x \in \Omega \,|\, x_1 = t\}$, that is, $u(t,x_2) = u(t,0)$ for $0 < x_2 < 1$, which implies $(\nabla u \cdot \nu)/|\nabla u| = \pm 1$ on $\partial R_t \cap \{x_1 = t\}$ in (3.4). Thus from (3.4), we have

$$0 = \int_{\partial R_t} \frac{a}{|\nabla u|}\frac{\partial u}{\partial \nu}\, ds = -1 \pm a(t).$$

By the knowledge of $a \geq 1$, we conclude that $a(t) = 1$ for all $t \in (0,1)$, which proves the necessity. ∎

4. Uniqueness of edge in a modified system. In order not to go astray from the main point of MREIT, we must focus on the final goal of MREIT, which aims to reconstruct the conductivity image σ. In section 3, we have observed that the model (3.2) may have infinitely many solutions u, and so has infinitely many distinct conductivity images $\sigma = a/|\nabla u|$. Thus, the model (3.2) is not appropriate for making a reconstruction algorithm. This is the main reason why the modified system (1.6) was introduced in [7] for the reconstruction algorithm that has been successfully demonstrated to provide accurate high-resolution conductivity images.

Although in numerical simulations in [7] the system (1.6) seems to have uniqueness, we were not able to prove the uniqueness rigorously as of this writing, but we could prove a practically useful uniqueness result which guarantees the unique detection of the edges of the conductivity image. This means that the system (1.6) uniquely determines the interface where the conductivity distribution σ is discontinuous.

As discussed at the beginning of section 3, if we assume that (u_1, u_2) is a practically acceptable solution of the system (1.6), we may impose the assumption

$$\sigma = \frac{a_1}{|\nabla u_1|} = \frac{a_2}{|\nabla u_2|} \in \Sigma,$$

which is the two-measurement analogue of the assumption (3.1). By plugging this assumption into the system (1.6), we have the following system, which will be considered in this section:

$$\begin{aligned} \nabla\cdot\left(\frac{a_j}{|\nabla u_j|}\nabla u_j\right) &= 0 \quad \text{in } \Omega, \\ \frac{a_1}{|\nabla u_1|} &= \frac{a_2}{|\nabla u_2|} \in \Sigma, \\ \frac{a_j}{|\nabla u_j|}\frac{\partial u_j}{\partial \nu} &= g_j \quad \text{on } \partial\Omega, \\ \int_{\partial\Omega} u_j\, ds &= 0 \end{aligned} \tag{4.1}$$

for $j = 1, 2$. For the uniqueness of (4.1), we need to choose an appropriate pair of current patterns g_1 and g_2 to have

$$|\nabla u_1(x) \times \nabla u_2(x)| > 0 \quad \text{for all } x \in \Omega. \tag{4.2}$$

In practice, each current g_j $(j = 1, 2)$ is applied through one pair of electrodes attached at points $P_j, Q_j \in \partial\Omega$. Here, the points P_1, P_2, Q_1, and Q_2 are situated along the boundary $\partial\Omega$ in this order and separated by a distance greater than 2ϵ. (See [7].) Hence we can assume, as in (1.2), the current g_j is approximated by

$$g_j(x) = \begin{cases} +\frac{I}{2\epsilon} & \text{on } \{|x - P_j| < \epsilon\} \cap \partial\Omega, \\ -\frac{I}{2\epsilon} & \text{on } \{|x - Q_j| < \epsilon\} \cap \partial\Omega, \\ 0 & \text{otherwise,} \end{cases} \tag{4.3}$$

where I is the current sent to both electrodes at P_j and Q_j, and 2ϵ is the width of each electrode. With these currents g_1 and g_2 as the Neumann data, from (2.5) we can easily see that the solution $(u_1, u_2) \in H^1(\Omega) \times H^1(\Omega)$ to the nonlinear system (4.1) satisfies

$$\nabla u_j(x) \neq 0 \quad \text{for all } x \in \Omega, \quad j = 1, 2.$$

More generally, in this case we can prove that (4.2) holds as the following lemma.

LEMMA 4.1. *Suppose that $(u_1, u_2) \in H^1(\Omega) \times H^1(\Omega)$ is a solution to the nonlinear system* (4.1) *with the Neumann data g_1 and g_2 defined in* (4.3). *Then we have*

$$|\nabla u_1(x) \times \nabla u_2(x)| > 0 \quad \textit{for all } x \in \Omega.$$

Proof. To derive a contradiction, suppose that there exists a point $\xi \in \Omega$ such that

$$|\nabla u_1(\xi) \times \nabla u_2(\xi)| = 0.$$

Then there exists a nonzero vector $(c_1, c_2) \in \mathbb{R}^2$ so that $c_1 \nabla u_1(\xi) + c_2 \nabla u_2(\xi) = 0$. Consider the function $w := c_1 u_1 + c_2 u_2$, which satisfies $\nabla w(\xi) = 0$ and

$$\nabla \cdot \left(\frac{a_1}{|\nabla u_1|} \nabla w \right) = 0 \quad \text{in } \Omega,$$
$$\frac{a_1}{|\nabla u_1|} \frac{\partial w}{\partial \nu} = \tilde{g} \quad \text{on } \partial\Omega, \quad \text{and} \quad \int_{\partial\Omega} w \, ds = 0,$$

where $\tilde{g} = c_1 g_1 + c_2 g_2$. By the assumption of

$$\frac{a_1(x)}{|\nabla u_1(x)|} = \frac{a_2(x)}{|\nabla u_2(x)|} \in \Sigma,$$

we may regard w as a solution to the classical Neumann problem (2.3) with the conductivity coefficient in the set Σ. Then all the properties in (2.4) hold for w.

On the other hand, the definition of g_j in (4.3) yields

$$\tilde{g}(x) = \begin{cases} +c_1 \frac{I}{2\epsilon} & \text{on } \{|x - P_1| < \epsilon\} \cap \partial\Omega, \\ +c_2 \frac{I}{2\epsilon} & \text{on } \{|x - P_2| < \epsilon\} \cap \partial\Omega, \\ -c_1 \frac{I}{2\epsilon} & \text{on } \{|x - Q_1| < \epsilon\} \cap \partial\Omega, \\ -c_2 \frac{I}{2\epsilon} & \text{on } \{|x - Q_2| < \epsilon\} \cap \partial\Omega, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, by the ordering of the points P_1, P_2, Q_1, and Q_2, we easily see that for any nonzero vector (c_1, c_2), $\tilde{g} \neq 0$ and there exist two disjoint arcs Γ^+ and Γ^- contained in $\partial\Omega$ such that

$$\Gamma^+ \cup \Gamma^- = \partial\Omega, \quad \text{and} \quad \Gamma^+ \subset \{\tilde{g} \geq 0\}, \quad \Gamma^- \subset \{\tilde{g} \leq 0\}. \tag{4.4}$$

Therefore, it follows from (2.5) that $\nabla w(x) \neq 0$ for all $x \in \Omega$. In particular, $\nabla w(\xi) \neq 0$, and hence it is a contradiction. This completes the proof. ∎

For the sake of clarity, we will give in the following remark more detailed proof for the reason why the property (4.4) of nonzero $\tilde{g}$ implies $\nabla w \neq 0$ in Ω, although it can also be found in [1, 2, 13].

Remark 4.2. Suppose that $\nabla w(\xi) = 0$; then by the maximum principle the level set $\{x \in \Omega \,|\, w(x) = w(\xi)\}$ divides Ω into more than four disjoint connected components $\Omega_1^{\pm}, \ldots, \Omega_m^{\pm}$ $(m \geq 2)$ such that (see Figure 4.1)

$$\bigcup_{k=1}^{m} \Omega_k^+ = \{x \in \Omega \,|\, w(x) > w(\xi)\} \quad \text{and} \quad \bigcup_{k=1}^{m} \Omega_k^- = \{x \in \Omega \,|\, w(x) < w(\xi)\}.$$

Applying the maximum principle again, we find that the boundary of each component $\Omega_k^{\pm}$ must occupy a portion $\gamma_k^{\pm}$ of $\partial\Omega$, that is, $\gamma_k^{\pm} := \partial\Omega_k^{\pm} \cap \partial\Omega \neq \emptyset$: if not, $\partial\Omega_k^{\pm}$ is a subset of the level curve $\{x \in \Omega \,|\, w(x) = w(\xi)\}$ and therefore by maximum principle w is the constant equal to $w(\xi)$ in $\Omega_k^{\pm}$. By the unique continuation, $\nabla w = 0$ in the whole domain Ω, and therefore $\tilde{g} = 0$, which is a contradiction.

From the maximum-minimum principle

$$\sup_{\Omega_k^+} w = \sup_{\partial\Omega_k^+} w = \sup_{\gamma_k^+} w \quad \text{and} \quad \inf_{\Omega_k^-} w = \inf_{\partial\Omega_k^-} w = \inf_{\gamma_k^-} w,$$

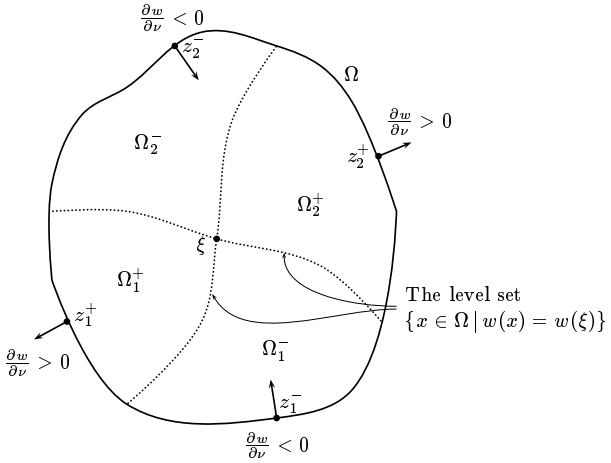


FIG. 4.1. *An illustration for Remark* 4.2 *when* $m = 2$.

there exist points $z_k^+ \in \gamma_k^+$ and $z_k^- \in \gamma_k^-$ so that

$$w(z_k^+) = \sup_{\Omega_k^+} w, \quad w(z_k^-) = \inf_{\Omega_k^-} w.$$

By Hopf's lemma, we have $\tilde{g}(z_k^+) > 0$ and $\tilde{g}(z_k^-) < 0$ for $k = 1, \ldots, m$. Since $m \geq 2$, $\tilde{g}$ cannot satisfy the property (4.4), which is a contradiction.

Lemma 4.1 tells us that two gradient vector fields ∇u_1 and ∇u_2 are neither vanishing nor parallel to each other at any points in Ω. Based on this fact, we can prove the following uniqueness result for the inverse problem with two measurements.

THEOREM 4.3. *Suppose that* $(u_1, u_2), (\tilde{u}_1, \tilde{u}_2) \in H^1(\Omega) \times H^1(\Omega)$ *are solutions to the nonlinear system* (4.1) *with the Neumann data* g_1 *and* g_2 *defined in* (4.3). *Then the edge of the conductivity image is uniquely determined by* (a_1, a_2) *in such a way that*

$$\left\{ x \in \Omega \;\middle|\; \frac{a_j}{|\nabla u_j|} \text{ is discontinuous at } x \right\} = \left\{ x \in \Omega \;\middle|\; \frac{a_j}{|\nabla \tilde{u}_j|} \text{ is discontinuous at } x \right\}.$$

Proof. Since (u_1, u_2) satisfies

$$\frac{a_1}{|\nabla u_1|} = \frac{a_2}{|\nabla u_2|} \in \Sigma,$$

there exist $\sigma_0 \in \mathcal{C}^\alpha(\bar{\Omega})$ and $\{(\sigma_k, D_k) \,|\, \sigma_k \in \mathcal{C}^\alpha(\bar{D}_k), \bar{D}_k \subset \Omega\}_{k=1}^M$ for some $M \in \mathbb{N}$, which satisfy

$$\frac{a_j}{|\nabla u_j|} = \sigma_0 + \sum_{k=1}^M \sigma_k \chi_{D_k} \in \Sigma. \tag{4.5}$$

Hence, from (2.1) we have

$$\sigma := \sigma_0 + \sum_{k=1}^M \sigma_k \chi_{D_k} \in \mathcal{C}^\alpha\left(\cup_{k=1}^M \bar{D}_k\right) \cap \mathcal{C}^\alpha\left(\Omega \setminus \cup_{k=1}^M D_k\right), \tag{4.6}$$

and u_j can be viewed as a solution of (2.3) when g is substituted by g_j. Thus, from (b) in (2.4) we get

$$\nabla u_j \in \mathcal{C}^\alpha\left(\cup_{k=1}^M \bar{D}_k\right) \cap \mathcal{C}^\alpha\left(\Omega \setminus \cup_{k=1}^M D_k\right). \tag{4.7}$$

From (4.5), we have $a_j(x) = \sigma(x)|\nabla u_j(x)|$, which implies that

$$a_j \in \mathcal{C}^\alpha\left(\cup_{k=1}^M \bar{D}_k\right) \cap \mathcal{C}^\alpha\left(\Omega \setminus \cup_{k=1}^M D_k\right) \qquad \text{for } j = 1, 2$$

by the aid of (4.6) and (4.7). Therefore, we get

$$A := \{x \in \Omega \,|\, a_1 \text{ or } a_2 \text{ is discontinuous at } x\} \subset \bigcup_{k=1}^M \partial D_k. \tag{4.8}$$

For the converse of (4.8), fix any $\xi \in \partial D_k$ for any $k = 1, \ldots, M$. It follows from Lemma 4.1 that either

$$\frac{\partial u_1^+}{\partial \tau}(\xi) \neq 0 \quad \text{or} \quad \frac{\partial u_2^+}{\partial \tau}(\xi) \neq 0, \tag{4.9}$$

where $u_j^+ := u_j|_{\Omega \setminus \bar{D}_k}$ for $j = 1, 2$, and $\partial/\partial\tau$ denotes the tangential derivative on ∂D_k. By the properties (c) and (d) in (2.4), we get

$$\sigma_0(\xi)\frac{\partial u_j^+}{\partial \nu}(\xi) = (\sigma_0(\xi) + \sigma_k(\xi))\frac{\partial u_j^-}{\partial \nu}(\xi) \quad \text{and} \quad \frac{\partial u_j^+}{\partial \tau}(\xi) = \frac{\partial u_j^-}{\partial \tau}(\xi),$$

where $u_j^- := u_j|_{D_k}$ and ν denotes the outward unit normal to ∂D_k. Considering $a_j = \sigma|\nabla u_j|$, a simple calculation yields that

$$|a_j^-(\xi)|^2 = |a_j^+(\xi)|^2 + \left((\sigma_0(\xi) + \sigma_k(\xi))^2 - (\sigma_0(\xi))^2\right)\left|\frac{\partial u_j^+}{\partial \tau}(\xi)\right|^2, \tag{4.10}$$

where $a_j^- := a_j|_{D_k}$ and $a_j^+ := a_j|_{\Omega \setminus \bar{D}_k}$. Since $\sigma_k(\xi) \neq 0$ by definition of Σ, by the aid of (4.9) the second term on the right-hand side of (4.10) is nonzero for either $j = 1$ or $j = 2$. Thus we show that a_1 or a_2 is discontinuous at ξ, and so $\xi \in A$. This proves that $\cup_{k=1}^M \partial D_k \subset A$. Hence, from (4.8) we conclude that $\cup_{k=1}^M \partial D_k = A$.

On the other hand, from (4.5) and (2.2), we can easily see that

$$\left\{x \in \Omega \;\middle|\; \frac{a_j}{|\nabla u_j|} \text{ is discontinuous at } x\right\} = \bigcup_{k=1}^M \partial D_k = A. \tag{4.11}$$

Because we have used only the fact that (u_1, u_2) is a solution to the nonlinear system (4.1), we can derive the same conclusion as (4.11) for $(\tilde{u}_1, \tilde{u}_2)$

$$\left\{x \in \Omega \;\middle|\; \frac{a_j}{|\nabla \tilde{u}_j|} \text{ is discontinuous at } x\right\} = \bigcup_{k=1}^{\tilde{M}} \partial \tilde{D}_k = A \tag{4.12}$$

for some mutually disjoint domains $\tilde{D}_k \subset \Omega$. Since the set A is completely determined by the data (a_1, a_2), the proof is completed by (4.11) and (4.12). □

Theorem 4.3 shows that the region where the conductivity distribution has jumps can be uniquely detected by the observation of discontinuities of the measured data

(a_1, a_2). In the following theorem, we show that the conductivity values as well as the unknown inclusions can be determined in a simple case when the conductivity distribution $\sigma \in \Sigma$ is known to be piecewise constant.

THEOREM 4.4. *Suppose that* $(u_1, u_2), (\tilde{u}_1, \tilde{u}_2) \in H^1(\Omega) \times H^1(\Omega)$ *are solutions to the nonlinear system* (4.1) *with the Neumann data* g_1 *and* g_2 *defined in* (4.3). *Suppose that* $\frac{a_j}{|\nabla u_j|}$ *and* $\frac{a_j}{|\nabla \tilde{u}_j|}$ *are piecewise constants, that is,*

$$\frac{a_j}{|\nabla u_j|} = 1 + \sum_{k=1}^{M} \mu_k \chi_{D_k} \quad \text{and} \quad \frac{a_j}{|\nabla \tilde{u}_j|} = 1 + \sum_{k=1}^{\tilde{M}} \tilde{\mu}_k \chi_{\tilde{D}_k}, \tag{4.13}$$

where $\mu_k, \tilde{\mu}_k$ *are nonzero constants satisfying* $-1 < \mu_k, \tilde{\mu}_k < \infty$. *Then* (u_1, u_2) *and* $(\tilde{u}_1, \tilde{u}_2)$ *are the same.*

Proof. First, we will prove that

$$\frac{a_j}{|\nabla u_j|} = \frac{a_j}{|\nabla \tilde{u}_j|}. \tag{4.14}$$

From (4.13), and (4.11), (4.12) in the proof of Theorem 4.3, the edge of the conductivity image is uniquely determined, that is, $M = \tilde{M}$ and $\bigcup_{k=1}^{M} D_k = \bigcup_{k=1}^{\tilde{M}} \tilde{D}_k$. Thus, for (4.14) it only remains to prove that $\mu_k = \tilde{\mu}_k$ for $k = 1, \ldots, M$. For this, it suffices to show that μ_k can be uniquely determined by the measured data (a_1, a_2) analogously as explained in the proof of Theorem 4.3. To be precise, μ_k will be shown to be determined by

$$\mu_k = \sqrt{1 + m_k} - 1, \quad k = 1, \ldots, M, \tag{4.15}$$

where the number m_k is defined by

$$m_k := \begin{cases} \max\limits_{\xi \in \partial D_k} \left\{ \left| \dfrac{a_1^-(\xi)}{a_1^+(\xi)} \right|^2 - 1 \right\} & \text{if } a_1^- \geq a_1^+ \text{ on } \partial D_k, \\ \\ \min\limits_{\xi \in \partial D_k} \left\{ \left| \dfrac{a_1^-(\xi)}{a_1^+(\xi)} \right|^2 - 1 \right\} & \text{if } a_1^- \leq a_1^+ \text{ on } \partial D_k. \end{cases} \tag{4.16}$$

Here, $a_1^- := a_1|_{D_k}$ and $a_1^+ := a_1|_{\Omega \setminus \bar{D}_k}$.

From (4.13), we have $a_1^+ = |\nabla u_1^+|$ on ∂D_k, and thus it follows that

$$\left| \frac{a_1^-(\xi)}{a_1^+(\xi)} \right|^2 - 1 = \frac{|\, a_1^-(\xi)|^2 - |\, a_1^+(\xi)|^2}{|\nabla u_1^+(\xi)|^2}, \quad \xi \in \partial D_k. \tag{4.17}$$

By the aid of (4.10) (in our case, $\sigma_0(\xi) = 1$ and $\sigma_k(\xi) = \mu_k$), we easily observe that either $a_1^- \geq a_1^+$ or $a_1^- \leq a_1^+$ on ∂D_k. In the case in which $a_1^- \geq a_1^+$, from (4.17) and (4.10) we have

$$\left| \frac{a_1^-(\xi)}{a_1^+(\xi)} \right|^2 - 1 \leq \frac{|\, a_1^-(\xi)|^2 - |\, a_1^+(\xi)|^2}{\left|\, \partial u_1^+ / \partial \tau(\xi) \,\right|^2} = \mu_k(\mu_k + 2)$$

for all $\xi \in \partial D_k$. In the case in which $a_1^- \leq a_1^+$, we get a similar result given by

$$\left| \frac{a_1^-(\xi)}{a_1^+(\xi)} \right|^2 - 1 \geq \mu_k(\mu_k + 2).$$

Now we will find the optimizer $z \in \partial D_k$ of (4.16). Applying the divergence theorem on $\Omega \setminus \bar{D}_k$, we get

$$0 = \int_{\partial\Omega} \frac{a_1}{|\nabla u_1|} \frac{\partial u_1}{\partial \nu}\, ds - \int_{\partial D_k} \frac{\partial u_1^+}{\partial \nu}\, ds = -\int_{\partial D_k} \frac{\partial u_1^+}{\partial \nu}\, ds,$$

noting that u_1 belongs to $\mathcal{C}^{1,\alpha}(\Omega \setminus \cup_{k=1}^M D_k)$ from (2.4) and $a_1^+/|\nabla u_1^+| = 1$ on ∂D_k. Hence there exists a point $z \in \partial D_k$ satisfying $\partial u_1^+/\partial \nu(z) = 0$, and by Lemma 4.1 we have

$$\left| \frac{\partial u_1^+}{\partial \tau}(z) \right| = |\nabla u_1^+(z)| > 0. \tag{4.18}$$

From (4.17), (4.18), and the jump relation (4.10), we obtain

$$\left| \frac{a_1^-(z)}{a_1^+(z)} \right|^2 - 1 = \mu_k(\mu_k + 2),$$

which implies that the point $z \in \partial D_k$ is the optimizer of (4.16). Thus it is clear that the number m_k defined in (4.16) is given by $m_k = \mu_k(\mu_k + 2) > -1$ because $\mu_k > -1$. Therefore we conclude that $\mu_k = \sqrt{1 + m_k} - 1$, which proves (4.15) and hence (4.14).

Finally, from (4.13) and (4.14) we see that u_j and $\tilde{u}_j$ can be viewed as the solutions of (2.3) when g is substituted by g_j and $\sigma := 1 + \sum_{k=1}^M \mu_k \chi_{D_k}$, since both (u_1, u_2) and $(\tilde{u}_1, \tilde{u}_2)$ are solutions to the nonlinear system (4.1). Hence by the uniqueness of the classical Neumann problem (2.3), we verify that $u_1 = \tilde{u}_1$ and $u_2 = \tilde{u}_2$, which completes the proof. ∎

5. Conclusion and numerical examples. A new reconstruction algorithm, the so-called *J-substitution algorithm*, was presented in [7] without uniqueness proofs to provide an impressively high-resolution conductivity image σ in simulations based on internal current density a obtained from the MRI system. For this algorithm, two different internal current densities a_1 and a_2 induced by two different applied currents g_1 and g_2 defined in (4.3) were used. In this paper, Theorem 4.3 has proved the uniqueness of the edge detection for piecewise continuous conductivities, and Theorem 4.4 has shown that a piecewise constant conductivity distribution can be completely reconstructed from a_1 and a_2.

On the other hand, it is worth investigating whether one could recover the conductivity distribution with only one internal current density, which means equivalently whether the nonlinear Neumann boundary value problem (3.2) could be solved uniquely. Theorem 3.1 has given a negative answer to this question.

In this section, we will present a numerically obtained example of nonuniqueness with one measurement which has been discussed in section 3. Suppose that Figure 5.1 represents an internal current density $a(x)$ on a cross-section $\Omega = (-1, 1) \times (-1, 1)$ of the human body induced by the applying the current

$$g(x) = \begin{cases} 1 & \text{if } x_1 = 1, \\ -1 & \text{if } x_1 = -1, \\ 0 & \text{otherwise,} \end{cases} \tag{5.1}$$

which can be viewed as an electrode attachment model in (1.2) when $P = (1, 0)$, $Q = (-1, 0)$, $I = 2$, and $\epsilon = 1$. We have numerically obtained this current density

$$a(x) := \sigma(x)|\nabla u(x)| \tag{5.2}$$

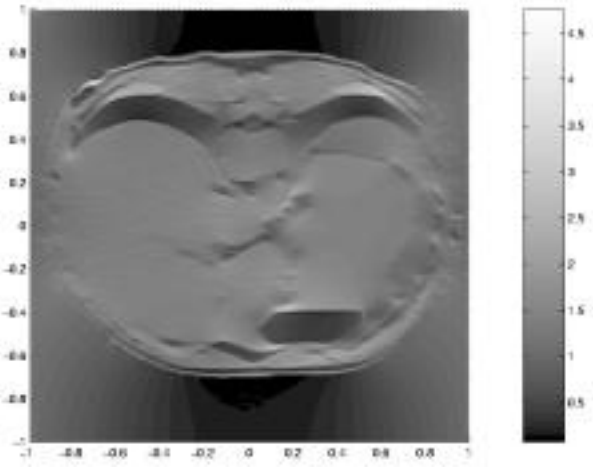
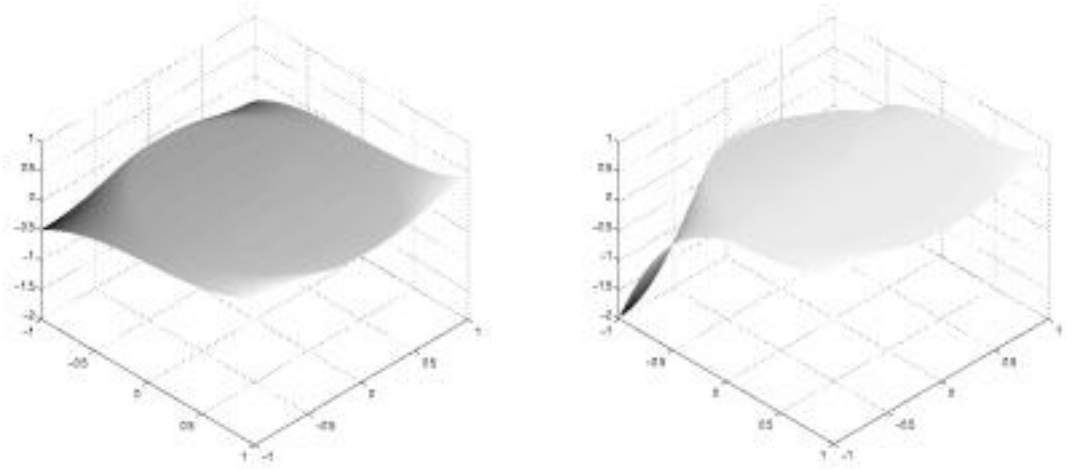


FIG. 5.1. *Simulated current density* $a(x)$.

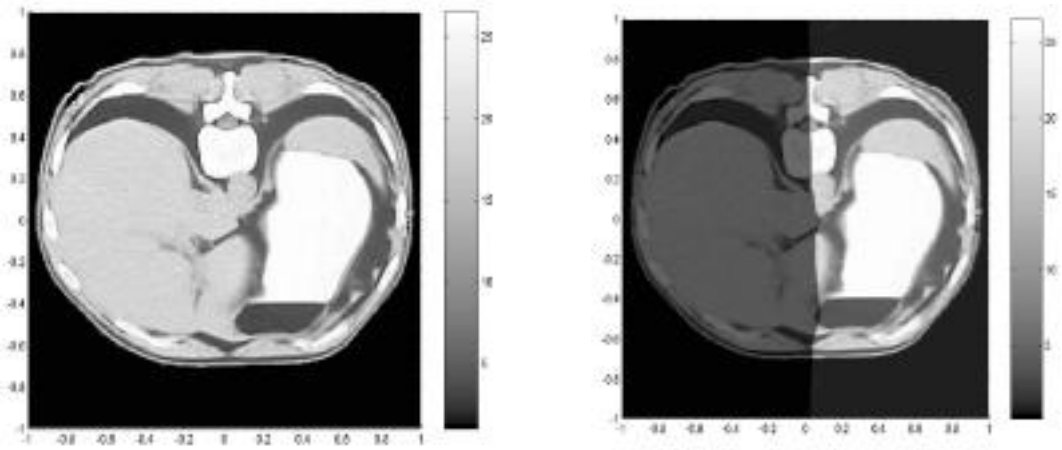


$u_1(x)$ $u_2(x)$

FIG. 5.2. *Two different solutions* u_1 *and* u_2 *to the problem* (3.2).

by assuming a conductivity distribution σ (in our experiment, σ is assumed to be σ_1 in Figure 5.3) and numerically solving the classical Neumann problem (2.3) with Neumann data g in (5.1) to calculate $|\nabla u(x)|$. As a numerical solver for (2.3), we have adopted the cell-centered finite difference scheme explained in [7]. In a real situation, the current density $a(x)$ is provided by a suitable MRI experiment called current density imaging [4, 6, 10, 11, 12, 15].

With this a and g, we can construct infinitely many solutions of the nonlinear Neumann boundary value problem (3.2) by virtue of Theorem 3.1. Here we present two different solutions u_1 and u_2, respectively given in Figure 5.2. Indeed, u_1 is equal to u that has been used to generate the simulated current density a in (5.2), and u_2 corresponds to $u_{t,\lambda}$ defined in the proof of Theorem 3.1 in the case in which $t = 0$



$\sigma_1(x)$ $\sigma_2(x)$

FIG. 5.3. *Two distinct conductivity images generated by* u_1 *and* u_2.

and $\lambda = 5$. These two different solutions yield two distinct conductivity images,

$$\sigma_1(x) = \frac{a(x)}{|\nabla u_1(x)|} \quad \text{and} \quad \sigma_2(x) = \frac{a(x)}{|\nabla u_2(x)|},$$

which are respectively shown in Figure 5.3. Hence, we conclude that only one internal current density information is insufficient for the unique determination of conductivity distributions.

REFERENCES

[1] G. ALESSANDRINI, V. ISAKOV, AND J. POWELL, *Local uniqueness in the inverse conductivity problem with one measurement*, Trans. Amer. Math. Soc., 347 (1995), pp. 3031–3041.

[2] G. ALESSANDRINI AND R. MAGNANINI, *The index of isolated critical points and solutions of elliptic equations in the plane*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 19 (1992), pp. 567–589.

[3] K.-S. CHENG, D. ISAACSON, J. C. NEWELL, AND D. G. GISSER, *Electrode model for electric current computed tomography*, IEEE Trans. Biomed. Engrg., 36 (1989), pp. 918–924.

[4] H. R. GAMBA AND D. T. DELPY, *Measurement of electrical current density distribution within the tissues of the head by magnetic resonance imaging*, Med. Biol. Engrg. Comp., 36 (1998), pp. 165–170.

[5] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[6] Y. Z. IDER AND L. T. MUFTULER, *Measurement of AC magnetic field distribution using magnetic resonance imaging*, IEEE Trans. Med. Imag., 16 (1997), pp. 617–622.

[7] O. KWON, E. J. WOO, J. R. YOON, AND J. K. SEO, *Magnetic resonance electrical impedance tomography (MREIT): Simulation study of J-substitution algorithm*, IEEE Trans. Biomed. Engrg., 49 (2002), pp. 160–167.

[8] O. A. LADYZHENSKAYA AND N. N. URAL'TSEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968, pp. 205–223.

[9] J. L. MUELLER, D. ISAACSON, AND J. C. NEWELL, *A reconstruction algorithm for electrical impedance tomography data collected on rectangular electrode array*, IEEE Trans. Biomed. Engrg., 46 (1999), pp. 1379–1386.

[10] G. C. SCOTT, M. L. G. JOY, R. L. ARMSTRONG, AND R. M. HENKELMAN, *Measurement of nonuniform current density by magnetic resonance*, IEEE Trans. Med. Imag., 10 (1991), pp. 362–374.

[11] G. C. Scott, M. L. G. Joy, R. L. Armstrong, and R. M. Henkelman, *Sensitivity of magnetic-resonance current density imaging*, J. Mag. Res., 97 (1992), pp. 235–254.
[12] G. C. Scott, M. L. G. Joy, R. L. Armstrong, and R. M. Henkelman, *Electromagnetic considerations for RF current density imaging*, IEEE Trans. Med. Imag., 14 (1995), pp. 515–524.
[13] J. K. Seo, *A uniqueness result on inverse conductivity problem with two measurements*, J. Fourier Anal. Appl., 2 (1996), pp. 227–235.
[14] J. G. Webster, ed., *Electrical Impedance Tomography*, Adam Hilger, Bristol, 1990.
[15] E. J. Woo, S. Y. Lee, and C. W. Mun, *Impedance tomography using internal current density distribution measured by nuclear magnetic resonance*, Proc. SPIE, 2299 (1994), pp. 377–385.

SIAM J. MATH. ANAL.
Vol. 34, No. 3, pp. 527–554

A FREE BOUNDARY PROBLEM FOR A HYPOPLASTIC MODEL OF PLANE SHEAR WAVES IN A FULLY SATURATED GRANULAR MATERIAL*

MICHAEL S. GORDON†

Abstract. A one-dimensional system describing small shearing disturbances in a semi-infinite, fully saturated granular medium is studied. The system is fully nonlinear as a result of the incrementally nonlinear constitutive law for the material. In particular, there are two different wave speeds corresponding to loading or unloading of the material. A free boundary problem for the boundary between loading and unloading regions is derived and solved globally. The solution is then applied to the investigation of two specific boundary value problems for the full system.

Key words. saturated granular material, hypoplastic flow rule, shear waves, free boundary problem, loading and unloading, characteristic analysis, functional equation

AMS subject classifications. 35R35, 39B22

PII. S0036141001390129

1. Introduction.

The following system of equations was derived by Osinov and Gudehus [7] as the key equations in a simplified model for plane shear waves in a saturated granular body:

$$\begin{aligned} \partial_t v &= \partial_x \sigma, \\ \partial_t \sigma &= a\partial_x v + b|\partial_x v|. \end{aligned} \tag{1.1}$$

Here a and b are constants satisfying $0 < b < a$. The dependent variables are v and σ; v is velocity and σ is a component of stress. Notice that, in regions where $\partial_x v$ (or, equivalently, $\partial_t \sigma$) does not change sign, the system (1.1) reduces to a linear wave equation with wave speed $\sqrt{a+b}$ or $\sqrt{a-b}$. Osinov and Gudehus derive the system (1.1) from a full three-dimensional system of equations for the deformation of a saturated granular material with a hypoplastic flow rule. (See [1], [4], [5], [6].) They linearize this system about a static state (i.e., zero strain rate) with constant stress tensor T^0; further, they assume that the incremental variables depend only on t and x_1 and that v_2 is the only nonzero component of velocity. The equations for $v = v_2$ and $\sigma = T_{12} - T^0_{12}$ (the perturbation of the shear stress) decouple from the other equations, leading to the system (1.1). We should point out that the constant b is positive because we assume that $T^0_{12} < 0$; the sign of b changes if $T^0_{12} > 0$. (See [7].) A consequence of $T^0_{12} < 0$ is that increasing σ decreases the magnitude of the total shear stress and thus unloads the material. Similarly, decreasing σ loads the material. Thus we will refer to regions where $\partial_t \sigma < 0$ (or, equivalently, $\partial_x v < 0$) as *loading* regions and those where $\partial_t \sigma > 0$ ($\partial_x v > 0$) as *unloading*.

A physical context for this model is shown in Figure 1.1. The figure shows a saturated granular material resting on an inclined solid mass. Plane shear waves described by (1.1) propagate in the direction perpendicular to the interface between the solid and the granular material, while the velocity vector is parallel to it. Boundary

*Received by the editors May 30, 2001; accepted for publication (in revised form) May 31, 2002; published electronically December 13, 2002.
http://www.siam.org/journals/sima/34-3/39012.html
†Department of Mathematics, State University of West Georgia, Carrollton, GA 30118 (sgordon@westga.edu).

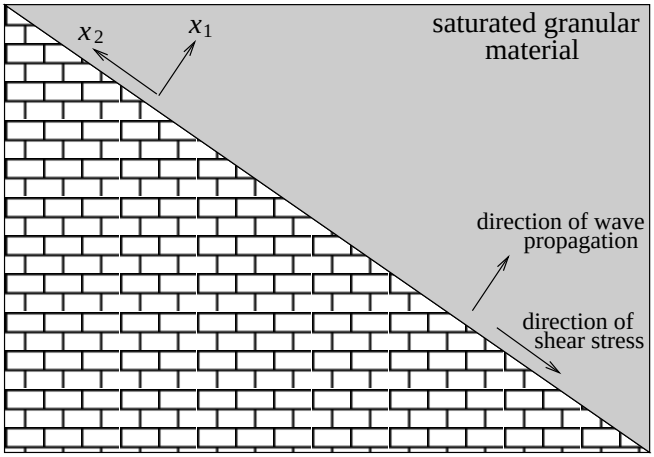


FIG. 1.1. *Physical setting for the model equations.*

disturbances could be created by waves propagating through the solid. A negative shear stress T_{12}^0 is created by the weight of the granular material.

We consider the system (1.1) in the quarter plane $x \geq 0$, $t \geq 0$ with initial data

$$v(x,0) = \sigma(x,0) = 0 \tag{1.2}$$

and stress-controlled boundary data

$$\sigma(0,t) = -\phi(t), \tag{1.3}$$

where $\phi(0) = 0$, $\lim_{t\to\infty} \phi(t) = 0$, ϕ is continuous, nonnegative, nondecreasing on $[0, t_s]$, and nonincreasing on $[t_s, \infty)$. (Two examples of such ϕ are shown in Figures 6.1 and 6.2.) Notice that (1.3) implies that σ decreases initially and then increases so that the resulting traveling pulse consists of a loading front followed by an unloading front. The form of the solution in the xt-plane (near $t = t_s$) is shown in Figure 1.2. Velocity and stress are zero in A_0 due to (1.2), A_1 is a loading region, and A_2 is an unloading region. Finding the solution in A_0 and A_1 is a straightforward application of the method of characteristics to (1.2) and (1.3). However, the solution in A_2 and the interface between A_1 and A_2 are interdependent, resulting in a free boundary problem. Gordon, Shearer, and Schaeffer [3] solve (1.1), (1.2), (1.3) with piecewise linear boundary data so that the resulting interface between loading and unloading regions is a straight line whose slope is the solution of a quadratic. In our case, this interface will be the solution of a difficult functional equation similar to the one solved in [2], and we find solutions with a similar iterative technique. The solution in [2] is local, and the iterations shrink the domain so that some effort is needed to show that it does not vanish to a point in the limit. In this work, we find a local solution and show that the iterations actually enlarge the domain, leading to a global solution.

In section 2, we use characteristic analysis to reduce the boundary value problem to an equation for the loading/unloading interface. In sections 3 and 4, we consider two cases: (i) ϕ has a corner at t_s and (ii) ϕ is smooth at t_s. The main challenge in the smooth case is finding a suitable function space, closed under our iterating operator, in which to seek a solution.

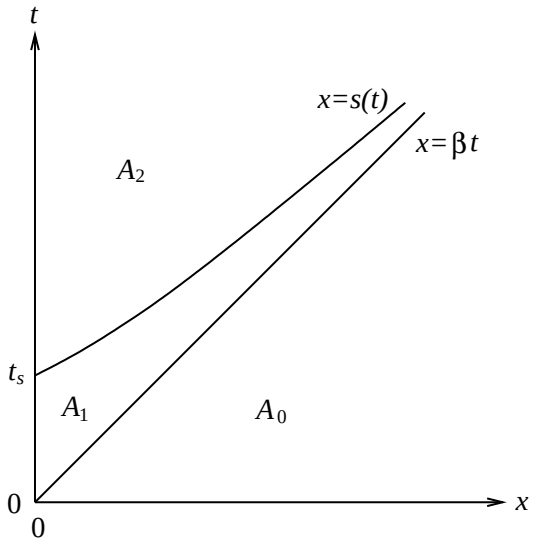


FIG. 1.2. *Form of the solution in xt-space.*

We had hoped to show in this work that the long-time solution of (1.1), (1.2), (1.3) was the same as was shown for a piecewise linear boundary pulse in [3], i.e., a decaying pulse which consists of a leading loading front and a trailing unloading front for all time. In fact, our derivation in section 2 of the equation for the loading/unloading interface assumes that our solution also has this form for all time. (See Figure 1.2.) However, if the solution begins to load somewhere in the unloading region after the loading/unloading interface has passed, it will invalidate that derivation. Unfortunately, this can happen for some choices of ϕ, as we show by example in section 6. In that section, we consider two examples of possible boundary data ϕ. We use analysis and numerical computations to show that, in one case, the solution of the interface equation does not lead to a global solution of (1.1), (1.2), (1.3), while in the other it does. In the second case, we show that the behavior of the solution is, as expected, qualitatively the same as for the stress-controlled problem in [3]. In section 5, we show that the solution of the interface equation leads to a solution of (1.1), (1.2), (1.3) that is at least locally valid.

2. Derivation of loading/unloading interface equation. In deriving an equation for the interface between loading and unloading regions, we will assume that the solution of (1.1), (1.2), (1.3) has the form shown in Figure 1.2; $v = \sigma = 0$ in $A_0 = \{(x,t) : x > \beta t\}$; $A_1 = \{(x,t) : s(t) < x < \beta t\}$ is a loading region, and $A_2 = \{(x,t) : x < s(t)\}$ is unloading. (We will show later that this assumption is at least locally valid and that σ and v are continuous across the loading/unloading interface $x = s\,(t)$.) This means that

$$\begin{aligned} \partial_t v &= \partial_x \sigma, \\ \partial_t \sigma &= \beta^2 \partial_x v \end{aligned} \tag{2.1}$$

in A_1, where $\beta = \sqrt{a-b}$ is the slow wave speed associated with loading, and

$$\begin{aligned} \partial_t v &= \partial_x \sigma, \\ \partial_t \sigma &= \alpha^2 \partial_x v \end{aligned} \tag{2.2}$$

in A_2, where $\alpha = \sqrt{a+b}$ is the fast wave speed associated with unloading. The curve $x = s(t)$ must satisfy the entropy condition discussed in [3]:

$$\beta \leq \frac{s(t+h) - s(t)}{h} \leq \alpha. \tag{2.3}$$

We will use characteristic analysis to derive an equation for the loading/unloading interface. First, notice that $v = \sigma = 0$ in A_0 because of (1.2) and the fact that every point in A_0 is connected to the x-axis by a pair of characteristics. Each point in A_1 is reached by one characteristic emanating from the x-axis and one from the t-axis, so the solution there is determined by (1.2) and (1.3). To see this, we rewrite (2.1) as

$$\begin{aligned} (\partial_t - \beta \partial_x)(\sigma + \beta v) &= 0, \\ (\partial_t + \beta \partial_x)(\sigma - \beta v) &= 0. \end{aligned} \tag{2.4}$$

Equations (2.4) and (1.2) imply that $\sigma + \beta v = 0$ in A_1, so

$$v = -\sigma/\beta \text{ in } A_1. \tag{2.5}$$

Combining (2.5) with (1.3), we have

$$v(0,t) = \phi(t)/\beta \quad \text{for } t \leq t_s.$$

This, (1.3), and (2.4) now imply that

$$(\sigma - \beta v)(0,t) = -2\phi(t) \quad \text{for } t \leq t_s. \tag{2.6}$$

Equation (2.6), combined with (2.4), determines $\sigma - \beta v$ on all of A_1; using (2.5), the entire solution σ, v is then determined on all of A_1. Notice that, if the loading/unloading interface $x = s(t)$ was known, characteristic analysis could then be used to find the entire solution on A_2, since two characteristics enter A_2 from the interface. Since this would also determine the known boundary condition $\sigma = \phi(t)$ on the t-axis for $t > t_s$, it seems reasonable to expect that we can set up an equation relating $s(t)$ to $\phi(t)$. Figure 2.1 shows how this will be accomplished. Referring to that figure, we let $(s(t), t)$ be a point on the interface and

$$\begin{aligned} t_\alpha &= t - s(t)/\alpha = \tilde{t} + s(\tilde{t})/\alpha, \\ t_\beta &= t - s(t)/\beta, \quad \tilde{t}_\beta = \tilde{t} - s(\tilde{t})/\beta. \end{aligned} \tag{2.7}$$

From (2.6) and (2.4), we have

$$(\sigma - \beta v)(s(t), t) = (\sigma - \beta v)(0, t_\beta) = -2\phi(t_\beta). \tag{2.8}$$

By (2.5),

$$(\sigma - \beta v)(s(t), t) = 2\sigma(s(t), t). \tag{2.9}$$

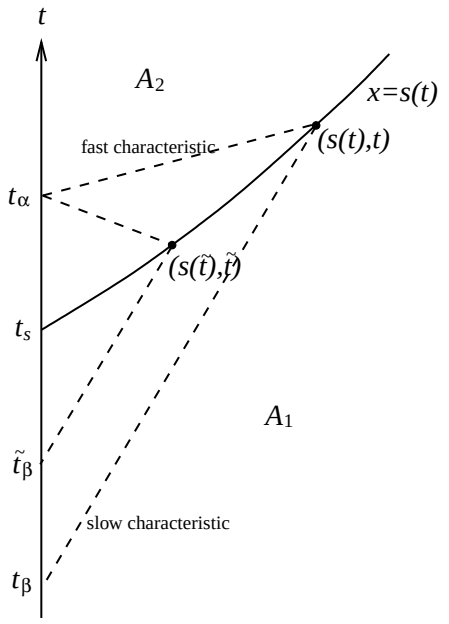


FIG. 2.1. *Derivation of the interface equation.*

Combining (2.5), (2.8), and (2.9), we have

$$\begin{aligned} \sigma(s(t),t) &= -\phi(t_\beta), \\ v(s(t),t) &= \phi(t_\beta)/\beta. \end{aligned} \tag{2.10}$$

Similarly,

$$\begin{aligned} \sigma(s(\tilde{t}),\tilde{t}) &= -\phi(\tilde{t}_\beta), \\ v(s(\tilde{t}),\tilde{t}) &= \phi(\tilde{t}_\beta)/\beta. \end{aligned} \tag{2.11}$$

Notice that (2.2) can be written as (2.4) with α in place of β:

$$\begin{aligned} (\partial_t - \alpha\partial_x)(\sigma + \alpha v) &= 0, \\ (\partial_t + \alpha\partial_x)(\sigma - \alpha v) &= 0. \end{aligned} \tag{2.12}$$

This implies

$$\begin{aligned} (\sigma - \alpha v)(s(t),t) &= (\sigma - \alpha v)(0,t_\alpha) = -\phi(t_\alpha) - \alpha v(0,t_\alpha), \\ (\sigma + \alpha v)(s(\tilde{t}),\tilde{t}) &= (\sigma + \alpha v)(0,t_\alpha) = -\phi(t_\alpha) + \alpha v(0,t_\alpha). \end{aligned}$$

Adding these gives

$$(\sigma - \alpha v)(s(t),t) + (\sigma + \alpha v)(s(\tilde{t}),\tilde{t}) = -2\phi(t_\alpha). \tag{2.13}$$

Notice that (2.10) and (2.12) imply

$$(\sigma - \alpha v)(s(t),t) + (\sigma + \alpha v)(s(\tilde{t}),\tilde{t}) = -(\alpha/\beta + 1)\,\phi(t_\beta) + (\alpha/\beta - 1)\,\phi(\tilde{t}_\beta). \tag{2.14}$$

Combining (2.13) and (2.14), we have the equation we seek:

$$-(\alpha/\beta+1)\,\phi(t_\beta)+(\alpha/\beta-1)\,\phi(\tilde{t}_\beta)=-2\phi(t_\alpha).$$

For simplicity, we rewrite this equation as

$$\mu\phi(t_\beta)-\phi(\tilde{t}_\beta)=(\mu-1)\phi(t_\alpha), \tag{2.15}$$

where $\mu=(\alpha+\beta)/(\alpha-\beta)>1$.

It becomes easier to see (2.15) in terms of the unknown interface if we make a change in coordinates so that the characteristics in A_2 become coordinate directions: $\xi=t-x/\alpha$, $\zeta=t+x/\alpha$. We let $\zeta=\rho(\xi)$ be the loading/unloading interface in the new variables, where ρ is defined implicitly in terms of s:

$$\rho\left(t-s\left(t\right)/\alpha\right)=t+s\left(t\right)/\alpha \quad \text{for } t>t_s. \tag{2.16}$$

The point $(t,s(t))$ becomes $(\xi,\rho(\xi))$ in the new coordinates. We let $\tilde{\xi}=\tilde{t}-s(\tilde{t})/\alpha$; then $(\tilde{t},s(\tilde{t}))$ becomes $(\tilde{\xi},\rho(\tilde{\xi}))$ in the new coordinates. Notice that (2.16) implies

$$\xi=t-s(t)/\alpha=t_\alpha \ \text{ by (2.7)}, \tag{2.17}$$

$$\rho\left(\xi\right)=t+s\left(t\right)/\alpha, \tag{2.18}$$

$$\rho(\tilde{\xi})=\tilde{t}+s(\tilde{t})/\alpha=t_\alpha \ \text{ by (2.7)}. \tag{2.19}$$

We solve (2.17) and (2.18) for t and $s\left(t\right)$ to get

$$\begin{aligned} t&=\left(\rho\left(\xi\right)+\xi\right)/2,\\ s\left(t\right)&=\alpha\left(\rho\left(\xi\right)-\xi\right)/2. \end{aligned} \tag{2.20}$$

We use (2.20) to rewrite t_β in terms of the new variables:

$$t_\beta=t-s(t)/\beta=\left(\rho\left(\xi\right)+\xi\right)/2-\frac{\alpha}{2\beta}\left(\rho\left(\xi\right)-\xi\right)=\frac{\mu\xi-\rho(\xi)}{\mu-1}. \tag{2.21}$$

A similar calculation with $\tilde{t}$ and $\tilde{\xi}$ in place of t and ξ gives

$$\tilde{t}_\beta=\tilde{t}-s(\tilde{t})/\beta=\frac{\mu\tilde{\xi}-\rho(\tilde{\xi})}{\mu-1}. \tag{2.22}$$

Notice that (2.17) and (2.19) imply

$$\xi=\rho(\tilde{\xi})\Rightarrow\tilde{\xi}=\rho^{-1}(\xi),$$

provided that ρ is invertible. Combining this with (2.22), we have

$$\tilde{t}_\beta=\frac{\mu\rho^{-1}(\xi)-\xi}{\mu-1}. \tag{2.23}$$

Using (2.17), (2.21), and (2.23) to make appropriate substitutions into (2.15), we now have

$$\mu\phi\left(\frac{\mu\xi-\rho(\xi)}{\mu-1}\right)-\phi\left(\frac{\mu\rho^{-1}(\xi)-\xi}{\mu-1}\right)=(\mu-1)\phi(\xi). \tag{2.24}$$

We have reduced the boundary value problem to solving (2.24) for the loading/unloading interface $\zeta = \rho(\xi)$. We seek a solution ρ, defined on $[t_s, \infty)$, satisfying the entropy condition (2.3) or, equivalently,

$$D_\ell \rho(\xi) \geq \mu \tag{2.25}$$

for all $\xi \geq t_s$, where

$$D_\ell \rho(\xi) = \liminf_{h \to 0^+} \frac{\rho(\xi + h) - \rho(\xi)}{h}$$

is the lower Dini derivative. The upper Dini derivative is defined by

$$D_u \rho(\xi) = \limsup_{h \to 0^+} \frac{\rho(\xi + h) - \rho(\xi)}{h}.$$

We will also show that the loading/unloading interface $x = s\,(t)$ approaches but never intersects the leading edge of the front $x = \beta t$, or, equivalently,

$$\rho(\xi) < \mu\xi \tag{2.26}$$

for all $\xi \geq t_s$, and

$$\lim_{\xi \to \infty} (\mu\xi - \rho(\xi)) = 0. \tag{2.27}$$

Notice that (2.27) and (2.25) imply that

$$\lim_{\xi \to \infty} D_\ell \rho(\xi) = \mu. \tag{2.28}$$

3. Corner case. In this section, we solve (2.24), assuming that there is a jump in the derivative of ϕ at t_s. More precisely, we let $\phi_1 = \phi|_{[0,t_s]}$ and $\phi_2 = \phi|_{[t_s,\infty)}$. We assume that $\phi_1 \in C^1\,[t_s - \delta_0, t_s]$, $\phi_2 \in C^1[t_s, t_s + \delta_0]$ for some $\delta_0 > 0$, and

$$\phi_1' > 0 \text{ on } [t_s - \delta_0, t_s] \text{ and } \phi_2' \leq 0 \text{ on } [t_s, t_s + \delta_0]. \tag{3.1}$$

The solution of (2.24) will be obtained by an iterative procedure defined as follows. Given a continuous function ρ satisfying (2.25) and $\rho(t_s) = t_s$, let $\Psi(\rho)$ be the function satisfying

$$\mu\phi_1\left(\frac{\mu\xi - \Psi(\rho)(\xi)}{\mu - 1}\right) - \phi_1\left(\frac{\mu\rho^{-1}(\xi) - \xi}{\mu - 1}\right) = (\mu - 1)\phi_2(\xi)$$

or, equivalently,

$$\Psi(\rho)(\xi) = \mu\xi - (\mu - 1)\phi_1^{-1}\left[(1 - 1/\mu)\phi_2(\xi) + \frac{1}{\mu}\phi_1\left(\frac{\mu\rho^{-1}(\xi) - \xi}{\mu - 1}\right)\right]. \tag{3.2}$$

Notice that a function ρ which is a fixed point of Ψ is a solution of (2.24). We will show that $\Psi^n(\rho)$ converges to a fixed point of Ψ as $n \to \infty$.

Before we continue, however, we pause to give some motivation for the choice of Ψ. Consider the case, solved in [3], where the boundary data ϕ is piecewise linear. As mentioned in the introduction, the loading/unloading interface in this case is a

straight line. Let $\phi_1'(t) = \omega_1 > 0$, $\phi_2'(t) = -\omega_2 \leq 0$, and $\rho'(\xi) = \rho_0$. It is not difficult to show that (2.24) reduces to

$$\mu\omega_1\left(\frac{\mu-\rho_0}{\mu-1}\right) - \omega_1\left(\frac{\mu/\rho_0-1}{\mu-1}\right) = -(\mu-1)\omega_2, \tag{3.3}$$

which then reduces to a quadratic in ρ_0. We examine what happens when we employ the iterative method described above to (3.3). Ψ is now defined by

$$\begin{aligned}\Psi(\rho_0) &= \mu - \frac{\mu-1}{\omega_1}\left[-(1-1/\mu)\omega_2 + \frac{\omega_1}{\mu}\left(\frac{\mu/\rho_0-1}{\mu-1}\right)\right] \\ &= \mu + \frac{\omega_2(\mu-1)^2}{\omega_1\mu} + 1/\mu - 1/\rho_0.\end{aligned}$$

Notice that $\Psi : [\mu, k] \to [\mu, k]$, where

$$k = \mu + \frac{\omega_2(\mu-1)^2}{\omega_1\mu} + 1/\mu. \tag{3.4}$$

Suppose $\rho_0, \bar{\rho}_0 \in [\mu, k]$. Then

$$|\Psi(\bar{\rho}_0) - \Psi(\rho_0)| = |1/\rho_0 - 1/\bar{\rho}_0| = \frac{|\bar{\rho}_0 - \rho_0|}{\bar{\rho}_0\rho_0} \leq \frac{|\bar{\rho}_0 - \rho_0|}{\mu^2}.$$

This shows that Ψ is a contraction on $[\mu, k]$, and so $\Psi^n(\rho_0)$ converges to a solution of (3.3) for any $\rho_0 \in [\mu, k]$. This is the strategy we use in solving the general problem in this and the following section. In both cases, we will show that Ψ is a contraction in the supremum norm on a suitably chosen function space. In the corner case, the space chosen is analogous to the interval used above in the piecewise linear case, placing upper and lower bounds on $D_u\rho$ and $D_\ell\rho$. The choice is more difficult in the smooth case; more discussion precedes that section.

In this and the following section, we let

$$t_\beta(\xi) = \frac{\mu\xi - \Psi(\rho)(\xi)}{\mu-1} = \phi_1^{-1}\left[(1-1/\mu)\phi_2(\xi) + \frac{1}{\mu}\phi_1\left(\tilde{t}_\beta(\xi)\right)\right]. \tag{3.5}$$

$\tilde{t}_\beta(\xi)$ is as defined in (2.23). (See Figure 3.1.) The following lemma motivates the definition of the function space for the corner case.

LEMMA 3.1. *Suppose* $\rho \in C[t_s, t_s+\delta]$, $\delta > 0$, $\rho(t_s) = t_s$, $\rho(\xi) \leq \mu\xi$, *and* ρ *satisfies* (2.25). *Then* $\Psi(\rho) \in C[t_s, \rho(t_s+\delta)]$, $\Psi(\rho)(t_s) = t_s$, $\Psi(\rho)(\xi) \leq \mu\xi$, *and* $\Psi(\rho)$ *satisfies* (2.25).

Proof. Suppose $\rho \in C[t_s, t_s+\delta]$, $\rho(t_s) = t_s$, $\rho(\xi) \leq \mu\xi$, and ρ satisfies (2.25). It is clear from (3.2) that $\Psi(\rho)(t_s) = t_s$. It follows from $\rho(\xi) \leq \mu\xi$ and (2.25) that

$$\xi \leq \mu\rho^{-1}(\xi) \leq \mu t_s + (\xi - t_s) \Rightarrow 0 \leq \tilde{t}_\beta(\xi) = \frac{\mu\rho^{-1}(\xi) - \xi}{\mu-1} \leq t_s$$

for all $\xi \in [t_s, \rho(t_s+\delta)]$. Notice then that

$$0 \leq (1-1/\mu)\phi_2(\xi) + \phi_1\left(\tilde{t}_\beta(\xi)\right)/\mu \leq \phi(t_s)$$

for all $\xi \in [t_s, \rho(t_s+\delta)]$. It then follows from (3.2) that $\Psi(\rho) \in C[t_s, \rho(t_s+\delta)]$ and that $\Psi(\rho)(\xi) \leq \mu\xi$. It follows from (2.25) that $\tilde{t}_\beta(\xi)$ is nonincreasing, and so

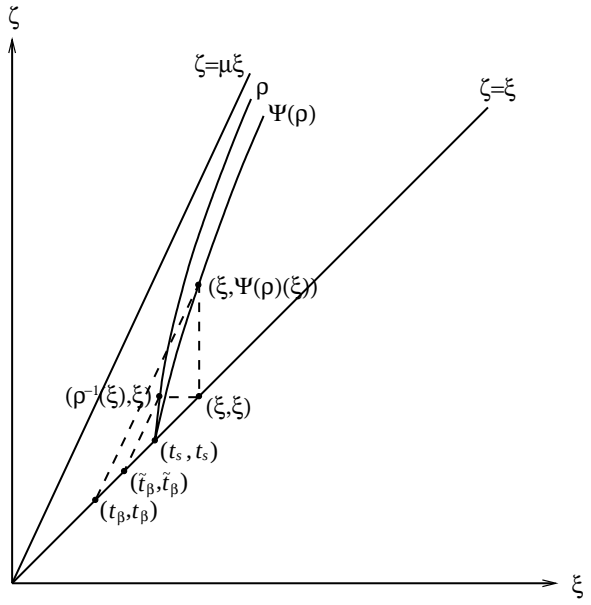


FIG. 3.1. *Iterative procedure for finding the interface.*

$\phi_1\left(\tilde{t}_\beta(\xi)\right)$ is nonincreasing. Since ϕ_2 is also nonincreasing, (3.5) implies that $t_\beta(\xi)$ is nonincreasing, and so

$$D_\ell \Psi(\rho)(\xi) = \mu - (\mu-1) D_u t_\beta(\xi) \geq \mu.$$

Hence $\Psi(\rho)$ satisfies (2.25). ∎

Let $K = \mu + (\mu-1)^2 K_1 K_2/\mu + K_1 K_2/\mu$, where $K_1 = \sup\left\{1/\phi_1'(t) : t_s - \delta_0 \leq t \leq t_s\right\}$ and $K_2 = \sup\left\{\left|\phi'(t)\right| : t_s - \delta_0 \leq t \leq t_s + \delta_0\right\}$. Notice that K_1 exists by (3.1). Define $\Gamma(\delta)$ to be the set of functions $\rho \in C[t_s, t_s+\delta]$ such that $\rho(t_s) = t_s$, $\rho(\xi) \leq \mu\xi$, ρ satisfies (2.25), and

$$D_u \rho(\xi) \leq K \tag{3.6}$$

for $\xi \in [t_s, t_s+\delta]$. We note that K is a generalization of the bound in (3.4).

We now show that $\Gamma(\delta)$ is closed under Ψ.

THEOREM 3.2. *$\Psi : \Gamma(\delta) \to \Gamma(\delta)$ for $\delta \leq \delta_0$.*

Proof. Suppose $\rho \in \Gamma(\delta_0)$. By Lemma 3.1, we need only show that $\Psi(\rho)$ satisfies (3.6). Differentiating (3.2), we have

$$\begin{aligned} D_u \Psi(\rho)(\xi) &= \mu - \frac{(\mu-1)^2}{\mu} \frac{\phi_2'(\xi)}{\phi_1'\left(t_\beta(\xi)\right)} + \frac{\phi_1'(\tilde{t}_\beta(\xi))}{\phi_1'\left(t_\beta(\xi)\right)} \left(\frac{1}{\mu} - \frac{1}{D_u \rho\left(\rho^{-1}(\xi)\right)}\right) \\ &\leq \mu + \frac{(\mu-1)^2}{\mu} K_1 K_2 + K_1 K_2/\mu = K. \end{aligned} \tag{3.7}$$

∎

THEOREM 3.3. *There is some $\delta_1 \in (0, \delta_0]$ such that Ψ is a contraction in the supremum norm on $\Gamma(\delta_1)$.*

Proof. Suppose $\rho_1, \rho_2 \in \Gamma(\delta_0)$. Let $\|\cdot\|_\delta$ denote the supremum norm on $\Gamma(\delta)$, and

let $t_{1,2}$, $\tilde{t}_{1,2}$ be defined as t_β, $\tilde{t}_\beta$ with $\rho_{1,2}$ in place of ρ. Then, by (3.2),

$$\begin{aligned}\Psi(\rho_1)(\xi) - \Psi(\rho_2)(\xi) &= (\mu-1)\left(t_2(\xi) - t_1(\xi)\right)\\ &= \frac{\mu-1}{\mu\phi_1'(\tau)}\left(\phi_1\left(\tilde{t}_2(\xi)\right) - \phi_1\left(\tilde{t}_1(\xi)\right)\right), \text{ where } \tau \text{ is between } t_1, t_2\\ &= \frac{(\mu-1)\,\phi_1'(\tilde{\tau})}{\mu\phi_1'(\tau)}\left(\tilde{t}_2(\xi) - \tilde{t}_1(\xi)\right), \text{ where } \tilde{\tau} \text{ is between } \tilde{t}_1, \tilde{t}_2\\ &= \frac{\phi_1'(\tilde{\tau})}{\phi_1'(\tau)}\left(\rho_2^{-1}(\xi) - \rho_1^{-1}(\xi)\right)\\ &= \frac{\phi_1'(\tilde{\tau})}{\phi_1'(\tau)}\frac{\rho_2^{-1}(\xi) - \rho_1^{-1}(\xi)}{\rho_1\left(\rho_2^{-1}(\xi)\right) - \rho_1\left(\rho_1^{-1}(\xi)\right)}\left(\rho_1\left(\rho_2^{-1}(\xi)\right) - \rho_2\left(\rho_2^{-1}(\xi)\right)\right).\end{aligned}$$

Thus, by (2.25),

$$|\Psi(\rho_1)(\xi) - \Psi(\rho_2)(\xi)| \le \frac{\phi_1'(\tilde{\tau})}{\mu\phi_1'(\tau)}\left|\rho_1\left(\rho_2^{-1}(\xi)\right) - \rho_2\left(\rho_2^{-1}(\xi)\right)\right|. \tag{3.8}$$

Choose $\kappa \in (1, \sqrt{\mu})$. By (3.1), we can choose $\delta_1 \in (0, \delta_0]$ so that

$$\nu/\kappa \le \phi_1'(t) \le \nu\kappa \quad \text{for } t \in [t_s, t_s + \delta_1(K-\mu)/(\mu-1)], \tag{3.9}$$

where $\nu = \phi_1'(t_s)$. By (3.6),

$$\rho(\xi) \le K(\xi - t_s) + t_s \tag{3.10}$$

for all $\rho \in \Gamma(\delta_1)$. Since $\tilde{t}_\beta$ is nonincreasing and $\rho(\xi) > \xi$, we have that $\tilde{t}_\beta(\xi) \ge \tilde{t}_\beta(\rho(\xi))$ which implies

$$t_s - \tilde{t}_\beta(\xi) \le t_s - \tilde{t}_\beta(\rho(\xi)) = t_s - \frac{\mu\xi - \rho(\xi)}{\mu-1}.$$

Combining this with (3.10), we have

$$t_s - \tilde{t}_\beta(\xi) \le \frac{K-\mu}{\mu-1}(\xi - t_s) \tag{3.11}$$

for all $\rho \in \Gamma(\delta_1)$. By Theorem 3.2, $\Psi(\rho) \in \Gamma(\delta_1)$, so $\Psi(\rho)$ satisfies (3.10). Thus, since

$$t_s - t_\beta(\xi) = t_s - \frac{\mu\xi - \Psi(\rho)(\xi)}{\mu-1},$$

t_β satisfies (3.11) for all $\rho \in \Gamma(\delta_1)$, and so $t_{1,2}$, $\tilde{t}_{1,2}$, τ, $\tilde{\tau}$ all satisfy (3.11) for $\xi \in [t_s, t_s + \delta_1]$. This, along with (3.9) and (3.8), implies that

$$\|\Psi(\rho_1) - \Psi(\rho_2)\|_{\delta_1} \le \frac{\nu\kappa}{\mu\nu/\kappa}\|\rho_1 - \rho_2\|_{\delta_1} = \frac{\kappa^2}{\mu}\|\rho_1 - \rho_2\|_{\delta_1},$$

from which the theorem follows. □

From Theorems 3.2 and 3.3 and the completeness of $\Gamma(\delta_1)$ in the supremum norm, we have the following theorem.

THEOREM 3.4. *$\Psi^n(\rho)$ converges to a unique solution of* (2.24) *in $\Gamma(\delta_1)$ for any $\rho \in \Gamma(\delta_1)$.*

We now show that the iterative procedure used to solve (2.24) locally actually leads to a global solution.

THEOREM 3.5. *There is a unique solution of* (2.24) *defined on* $[t_s, \infty)$ *satisfying* (2.25), (2.26), *and* (2.27).

Proof. Let $\rho_1 \in \Gamma(\delta_1)$ be a solution of (2.24), and let $\rho_n = \Psi(\rho_{n-1})$ for $n \geq 2$. Define δ_n by

$$t_s + \delta_n = \rho_{n-1}(t_s + \delta_{n-1}) \tag{3.12}$$

for $n \geq 2$. Notice that the proofs of Lemma 3.1 and Theorem 3.2 imply that $\rho_n \in \Gamma(\delta_n)$ for $n \geq 2$. By (2.25), $\delta_n \geq \mu\delta_{n-1}$, and so $\delta_n \to \infty$ as $n \to \infty$, and so ρ_n converges to a function ρ_* solving (2.24) on $[t_s, \infty)$ and satisfying (2.25).

We now show that ρ_* satisfies (2.26). Clearly there is some $\delta > 0$ such that $\rho_*(\xi) < \mu\xi$ for $\xi \in [t_s, t_s + \delta]$. This implies

$$\xi < \mu\rho_*^{-1}(\xi) \Rightarrow \tilde{t}_\beta(\xi) = \frac{\mu\rho_*^{-1}(\xi) - \xi}{\mu - 1} > 0$$

for all $\xi \in [t_s, \rho_*(t_s + \delta)]$. Thus

$$(1 - 1/\mu)\phi_2(\xi) + \frac{1}{\mu}\phi_1\left(\frac{\mu\rho_*^{-1}(\xi) - \xi}{\mu - 1}\right) > 0$$

for all $\xi \in [t_s, \rho_*(t_s + \delta)]$ since ϕ_1 is positive on $(0, t_s]$. Thus, by (3.2), $\rho_*(\xi) < \mu\xi$ for $\xi \in [t_s, \rho_*(t_s + \delta)]$. By (3.12), $\rho_*^n(t_s + \delta) = t_s + \delta_n \to \infty$ as $n \to \infty$, so $\rho_*(\xi) < \mu\xi$ for $\xi \in [t_s, \infty)$.

We now show that ρ_* satisfies (2.27). Equations (2.25) and (2.26) imply that $\mu\xi - \rho_*(\xi)$ is nonincreasing and bounded below by zero. Thus $\varepsilon = \lim_{\xi\to\infty}(\mu\xi - \rho_*(\xi)) \geq 0$. Letting $\xi \to \infty$ in (2.24), we have

$$\mu\phi_1\left(\frac{\varepsilon}{\mu - 1}\right) - \phi_1\left(\frac{\varepsilon}{\mu - 1}\right) = 0 \Rightarrow \phi_1\left(\frac{\varepsilon}{\mu - 1}\right) = 0 \Rightarrow \varepsilon = 0. \qquad \square$$

The following theorem shows that the solution ρ is differentiable if the boundary data ϕ is differentiable except at $t = t_s$.

THEOREM 3.6. *Suppose that* ϕ_1 *is differentiable on* $[0, t_s]$ *and* ϕ_2 *is differentiable on* $[t_s, \infty)$. *Then the solution* ρ *of* (2.24) *is differentiable on* $[t_s, \infty)$ *and*

$$\lim_{\xi\to\infty} \rho'(\xi) = \mu. \tag{3.13}$$

Proof. Let ρ be a solution of (2.24). Using (3.2) and differentiating, we have

$$\rho'(\xi) = \mu - \frac{(\mu - 1)^2}{\mu}\frac{\phi_2'(\xi)}{\phi_1'(t_\beta(\xi))} + \frac{\phi_1'(\tilde{t}_\beta(\xi))}{\phi_1'(t_\beta(\xi))}\left(\frac{1}{\mu} - \frac{1}{\rho'(\rho^{-1}(\xi))}\right), \tag{3.14}$$

which implies that if ρ is differentiable on $[t_s, t_s + \delta_1]$, then ρ is differentiable on $[t_s, \rho^n(t_s + \delta_1)]$, and hence on $[t_s, \infty)$. Thus we need only show that ρ is differentiable on $[t_s, t_s + \delta_1]$. Notice that $D_\ell\rho$ and $D_u\rho$ satisfy (3.14). Taking the difference, we have

$$\begin{aligned} D_u\rho(\xi) - D_\ell\rho(\xi) &= -\frac{\phi_1'(\tilde{t}_\beta(\xi))}{\phi_1'(t_\beta(\xi))}\left(\frac{1}{D_u\rho(\rho^{-1}(\xi))} - \frac{1}{D_\ell\rho(\rho^{-1}(\xi))}\right) \\ &= \frac{\phi_1'(\tilde{t}_\beta(\xi))}{\phi_1'(t_\beta(\xi))}\left(\frac{D_u\rho(\rho^{-1}(\xi)) - D_\ell\rho(\rho^{-1}(\xi))}{D_u\rho(\rho^{-1}(\xi))\, D_\ell\rho(\rho^{-1}(\xi))}\right). \end{aligned}$$

Combining this with (2.25) and (3.9), we have

$$|D_u\rho(\xi) - D_\ell\rho(\xi)| \le \frac{\kappa^2}{\mu^2}\left|D_u\rho\left(\rho^{-1}(\xi)\right) - D_\ell\rho\left(\rho^{-1}(\xi)\right)\right|$$

for $\xi \in [t_s, t_s + \delta_1]$. By induction,

$$|D_u\rho(\xi) - D_\ell\rho(\xi)| \le \frac{\kappa^{2n}}{\mu^{2n}}\left|D_u\rho\left(\rho^{-n}(\xi)\right) - D_\ell\rho\left(\rho^{-n}(\xi)\right)\right| \le \frac{\kappa^{2n}}{\mu^{2n}}K \to 0 \text{ as } n \to \infty,$$

and so $D_u\rho = D_\ell\rho$ on $[t_s, t_s + \delta_1]$. Equation (3.13) follows from (2.28). □

4. Smooth case. We now assume in place of (3.1) that ϕ is differentiable at t_s. Specifically, $\phi \in C^1[t_s - \delta_0, t_s + \delta_0]$ for some $\delta_0 > 0$ (so $\phi'(t_s) = 0$) and $\phi''(t_s)$ exists and is nonzero. Thus we may write

$$\phi(\xi) = \phi(t_s) - (\xi - t_s)^2(\lambda + \phi_0(\xi)), \tag{4.1}$$

where $\phi_0 \in C^1\left([t_s - \delta_0, t_s + \delta_0] - \{t_s\}\right)$, $\lim_{\xi\to t_s}\phi_0(\xi) = 0$, $\lim_{\xi\to t_s}(\xi - t_s)\,\phi_0'(\xi) = 0$, and $\lambda = -\frac{1}{2}\phi''(t_s) > 0$. We note that Lemma 3.1 still holds in this setting.

Before we begin, we remark on the difficulties of this case as compared to the corner case. The corner case relies on the Lipschitz continuity of ϕ_1 and its inverse to find a space of Lipschitz continuous functions which is complete and closed under our iterating operator. In the smooth case, ϕ_1^{-1} is not Lipschitz continuous, but we are still able to find a similar space of Lipschitz continuous functions by modifying the upper bound on difference quotients in (3.6). (See (4.2).) However, it is more difficult to show closure (Lemma 4.2) and requires a bound on the modulus of continuity of difference quotients of ρ at $\xi = t_s$ (see (4.3)), forcing another condition on the function space which must be preserved under Ψ (Lemma 4.1).

Let $\rho_0 = \mu - 1 + \sqrt{\mu^2 - \mu + 1} > \mu$. It will be apparent from the construction that follows that if ρ is a solution of (2.24) under the above assumptions, then $\rho'(t_s) = \rho_0$. Choose $\tilde{K} > \mu + 1/\mu + (\mu-1)^3/\left(\mu\rho_0 - \mu^2\right) > \rho_0$.

Define $\bar{\phi}_0(\xi) = \sup\{|\phi_0(\zeta)| : |\zeta - t_s| \le |\xi - t_s|\}$ and

$$\tilde{t}^*_\beta(\xi) = \frac{\mu t_s - \xi}{\mu - 1} = t_s - \frac{\xi - t_s}{\mu - 1},$$
$$t^*_\beta(\xi) = \phi_1^{-1}\left[(1 - 1/\mu)\phi_2(\xi) + \frac{1}{\mu}\phi_1\left(\tilde{t}^*_\beta(\xi)\right)\right].$$

Notice that $\tilde{t}^*_\beta$, t^*_β are lower bounds (independent of ρ) of $\tilde{t}_\beta$, t_β. Choose $c \in (1/\rho_0^3, 1)$ and define

$$\epsilon(h) = \frac{4(\rho_0 - \mu)\bar{\phi}_0(\tilde{t}^*_\beta(t_s + h))}{\lambda\mu\left(2\rho_0 - \mu\right)^2(1 - c)} + \frac{4(\rho_0 - \mu)\bar{\phi}_0(t^*_\beta(t_s + h))}{\lambda(1 - c)} + \frac{(\mu - 1)^3\bar{\phi}_0(t_s + h)}{\lambda\mu(\rho_0 - \mu)(1 - c)}.$$

Notice that ϵ is nondecreasing and $\lim_{h\to 0}\epsilon(h) = 0$.

Define $\Omega(\delta)$ to be the set of functions $\rho \in C[t_s, t_s + \delta]$ such that $\rho(t_s) = t_s$,

$$D_\ell\rho(\xi) \ge \mu, \quad D_u\rho(\xi) \le \tilde{K} \tag{4.2}$$

for $\xi \in [t_s, t_s + \delta]$, and

$$|D(\rho, h) - \rho_0| \le \epsilon(h) \tag{4.3}$$

for $h \in (0, \delta]$, where

$$D(\rho, h) = \frac{\rho(t_s + h) - \rho(t_s)}{h} = \frac{\rho(t_s + h) - t_s}{h}.$$

Notice that (4.3) and $\lim_{h \to 0} \epsilon(h) = 0$ imply that $\rho'(t_s) = \rho_0$ for all $\rho \in \Omega(\delta)$.

The following two lemmas, along with Lemma 3.1, establish the closure of Ω under Ψ.

LEMMA 4.1. *There is some $\delta_1 \in (0, \delta_0]$ such that $|D(\Psi(\rho), h) - \rho_0| \leq \epsilon(h)$ for all $\rho \in \Omega(\delta_1)$, $h \in (0, \delta_1]$.*

Proof. Let $\rho \in \Omega(\delta_0)$. Letting $\xi = t_s + h$ in (4.1) and substituting into (2.24), we have

$$\mu (t_\beta - t_s)^2 (\lambda + \phi_0(t_\beta)) - \left(\tilde{t}_\beta - t_s\right)^2 \left(\lambda + \phi_0\left(\tilde{t}_\beta\right)\right) = (\mu - 1) h^2 (\lambda + \phi_0(t_s + h)),$$

where $t_\beta = t_\beta(t_s + h)$, $\tilde{t}_\beta = \tilde{t}_\beta(t_s + h)$. Using (2.21) and (3.5), we have

$$\frac{\mu}{(\mu - 1)^2} (\Psi(\rho)(t_s + h) - t_s - h\mu)^2 (\lambda + \phi_0(t_\beta)) - \frac{1}{(\mu - 1)^2} \left(\mu\rho^{-1} - \mu t_s - h\right)^2 \left(\lambda + \phi_0\left(\tilde{t}_\beta\right)\right) = (\mu - 1) h^2 (\lambda + \phi_0(t_s + h))$$

which implies

$$\mu (D(\Psi(\rho), h) - \mu)^2 (\lambda + \phi_0(t_\beta)) - \left(1 - \mu D(\rho^{-1}, h)\right)^2 \left(\lambda + \phi_0\left(\tilde{t}_\beta\right)\right) = (\mu - 1)^3 (\lambda + \phi_0(t_s + h)).$$

Using the fact that $D(\rho^{-1}, h) = 1/D(\rho, \rho^{-1}(t_s + h) - t_s)$, this implies

$$\mu (D_1 - \mu)^2 (\lambda + \phi_0(t_\beta)) - (1 - \mu/D_2)^2 \left(\lambda + \phi_0\left(\tilde{t}_\beta\right)\right) = (\mu - 1)^3 (\lambda + \phi_0(t_s + h)), \tag{4.4}$$

where $D_1 = D(\Psi(\rho), h)$ and $D_2 = D(\rho, \rho^{-1}(t_s + h) - t_s)$. It is not hard to show that ρ_0 satisfies

$$\mu (\rho_0 - \mu)^2 - (1 - \mu/\rho_0)^2 = (\mu - 1)^3. \tag{4.5}$$

Multiplying (4.5) by λ and subtracting the result from (4.4), we have

$$\lambda\mu (D_1 - \rho_0)(D_1 + \rho_0 - 2\mu) + \mu (D_1 - \mu)^2 \phi_0(t_\beta) + \lambda\mu (1/\rho_0 - 1/D_2)(2 - \mu/\rho_0 - \mu/D_2) - (1 - \mu/D_2)^2 \phi_0\left(\tilde{t}_\beta\right) = (\mu - 1)^3 \phi_0(t_s + h),$$

and so

$$D_1 - \rho_0 = \frac{(D_2 - \rho_0)(\mu D_2 + \mu\rho_0 - 2\rho_0 D_2)}{\rho_0^2 D_2^2 (D_1 + \rho_0 - 2\mu)} + \frac{(1 - \mu/D_2)^2 \phi_0\left(\tilde{t}_\beta\right)}{\lambda\mu (D_1 + \rho_0 - 2\mu)} - \frac{(D_1 - \mu)^2 \phi_0(t_\beta)}{\lambda (D_1 + \rho_0 - 2\mu)} + \frac{(\mu - 1)^3 \phi_0(t_s + h)}{\lambda\mu (D_1 + \rho_0 - 2\mu)}.$$

Notice that

$$\frac{\mu D_2 + \mu\rho_0 - 2\rho_0 D_2}{\rho_0^2 D_2^2 (D_1 + \rho_0 - 2\mu)} \to -\frac{1}{\rho_0^3} \quad \text{as} \quad D_1, D_2 \to \rho_0.$$

Choose $\varepsilon \in (0, \rho_0 - \mu)$ such that

$$\left| \frac{\mu D_2 + \mu\rho_0 - 2\rho_0 D_2}{\rho_0^2 D_2^2 \left(D_1 + \rho_0 - 2\mu\right)} \right| < c \text{ when } |D_{1,2} - \rho_0| < \varepsilon. \tag{4.6}$$

Now, solving (4.4) for D_1, we have

$$D_1 = \mu + \left[\frac{(\mu-1)^3}{\mu} \cdot \frac{\lambda + \phi_0\left(t_s + h\right)}{\lambda + \phi_0\left(t_\beta\right)} + \frac{1}{\mu}\left(1 - \mu/D_2\right)^2 \cdot \frac{\lambda + \phi_0\left(\tilde{t}_\beta\right)}{\lambda + \phi_0\left(t_\beta\right)} \right]^{1/2}. \tag{4.7}$$

Equation (2.25) implies that $\rho^{-1}(t_s + h) - t_s \leq h/\mu < h$, and so, by (4.3),

$$|D_2 - \rho_0| \leq \epsilon(h). \tag{4.8}$$

Combining (4.7) and (4.8) we have upper and lower bounds on D_1 given by

$$\mu + \left[\frac{(\mu-1)^3}{\mu} \cdot \frac{\lambda \pm \bar{\phi}_0\left(t_s + h\right)}{\lambda \mp \bar{\phi}_0(t^*_\beta)} + \frac{1}{\mu}\left(1 - \frac{\mu}{\rho_0 \pm \epsilon(h)}\right)^2 \cdot \frac{\lambda \pm \bar{\phi}_0(\tilde{t}^*_\beta)}{\lambda \mp \bar{\phi}_0(t^*_\beta)} \right]^{1/2}, \tag{4.9}$$

where $t^*_\beta = t^*_\beta\left(t_s + h\right)$, $\tilde{t}^*_\beta = \tilde{t}^*_\beta\left(t_s + h\right)$. From (4.5) we have

$$\rho_0 = \mu + \left[\frac{(\mu-1)^3}{\mu} + \frac{1}{\mu}\left(1 - \frac{\mu}{\rho_0}\right)^2 \right]^{1/2}. \tag{4.10}$$

Equations (4.8), (4.9), and (4.10) imply that there is some $\delta_1 > 0$ such that $|D_{1,2} - \rho_0| < \varepsilon$ for all $\rho \in \Omega(\delta_1)$. Thus, by (4.6),

$$\begin{aligned} |D_1 - \rho_0| &\leq c\,|D_2 - \rho_0| + \frac{\left(1 - \mu/D_2\right)^2 \bar{\phi}_0\left(\tilde{t}_\beta\right)}{\lambda\mu\left(D_1 + \rho_0 - 2\mu\right)} + \frac{\left(D_1 - \mu\right)^2 \bar{\phi}_0\left(t_\beta\right)}{\lambda\left(D_1 + \rho_0 - 2\mu\right)} \\ &\qquad + \frac{(\mu-1)^3 \bar{\phi}_0\left(t_s + h\right)}{\lambda\mu\left(D_1 + \rho_0 - 2\mu\right)} \\ &\leq c\,|D_2 - \rho_0| + \frac{\left(1 - \mu/\left(\rho_0 + \varepsilon\right)\right)^2 \bar{\phi}_0(\tilde{t}^*_\beta)}{\lambda\mu\left(2\rho_0 - 2\mu - \varepsilon\right)} + \frac{\left(\rho_0 - \mu + \varepsilon\right)^2 \bar{\phi}_0(t^*_\beta)}{\lambda\left(2\rho_0 - 2\mu - \varepsilon\right)} \\ &\qquad + \frac{(\mu-1)^3 \bar{\phi}_0\left(t_s + h\right)}{\lambda\mu\left(2\rho_0 - 2\mu - \varepsilon\right)}. \end{aligned}$$

Using (4.8) and the fact that $\varepsilon < \rho_0 - \mu$, we have

$$\begin{aligned} |D_1 - \rho_0| &\leq c\epsilon(h) + \frac{4\left(\rho_0 - \mu\right)\bar{\phi}_0(\tilde{t}^*_\beta)}{\lambda\mu\left(2\rho_0 - \mu\right)^2} + \frac{4\left(\rho_0 - \mu\right)\bar{\phi}_0(t^*_\beta)}{\lambda} + \frac{(\mu-1)^3 \bar{\phi}_0\left(t_s + h\right)}{\lambda\mu\left(\rho_0 - \mu\right)} \\ &= \epsilon(h) \end{aligned}$$

for all $\rho \in \Omega(\delta_1)$. □

LEMMA 4.2. *There is some $\delta_2 \in (0, \delta_1]$ such that $D_u\Psi\left(\rho\right)\left(\xi\right) \leq \tilde{K}$ for all $\rho \in \Omega(\delta_2)$, $\xi \in [t_s, t_s + \delta_2]$.*

Proof. Let $\rho \in \Omega(\delta_1)$. From (3.7) we have

$$D_u\Psi(\rho)(\xi) \leq \mu + \frac{(\mu-1)^2}{\mu} \cdot \frac{\left|\phi_2'(\xi)\right|}{\phi_1'\left(t_\beta(\xi)\right)} + \frac{\phi_1'(\tilde{t}_\beta(\xi))}{\mu\phi_1'\left(t_\beta(\xi)\right)}. \tag{4.11}$$

Notice that

$$\tilde{t}_\beta(\xi) = t_s - \frac{\xi - t_s}{\mu - 1}\left(1 - \mu D(\rho^{-1}, \xi - t_s)\right) = t_s - \frac{\xi - t_s}{\mu - 1}\left(1 - \mu / D(\rho, \rho^{-1}(\xi) - t_s)\right),$$

so, by (4.8),

$$\tilde{t}_\beta(\xi) \geq t_s - \frac{\xi - t_s}{\mu - 1}\left(1 - \frac{\mu}{\rho_0 + \epsilon(\xi - t_s)}\right). \tag{4.12}$$

Also, from (3.5),

$$\begin{aligned} t_\beta(\xi) &= t_s - \frac{\xi - t_s}{\mu - 1}\left(D(\Psi(\rho), \xi - t_s) - \mu\right) \\ &\leq t_s - \frac{\xi - t_s}{\mu - 1}\left(\rho_0 - \epsilon(\xi - t_s) - \mu\right) \end{aligned} \tag{4.13}$$

by Lemma 4.1. Given (4.12), (4.13), and

$$1 - \mu/\rho_0 < \rho_0 - \mu, \tag{4.14}$$

we can choose $\delta_3 \in (0, \delta_1]$ such that

$$t_\beta(\xi) \leq \tilde{t}_\beta(\xi) \text{ for all } \rho \in \Omega(\delta_3),\ \xi \in [t_s, t_s + \delta_3]. \tag{4.15}$$

Choose $\delta_4 \in (0, \delta_3]$ such that ϕ_1' is decreasing on $[t_\beta^*(t_s + \delta_4), t_s]$. Then, by (4.15), we have

$$\frac{\phi_1'\left(\tilde{t}_\beta(\xi)\right)}{\phi_1'\left(t_\beta(\xi)\right)} \leq 1 \text{ for all } \rho \in \Omega(\delta_4),\ \xi \in [t_s, t_s + \delta_4]. \tag{4.16}$$

Differentiating (4.1), we have

$$\begin{aligned} \frac{\left|\phi_2'(\xi)\right|}{\phi_1'(t_\beta)} &= \frac{(\xi - t_s)\left(2\lambda + 2\phi_0(\xi) + (\xi - t_s)\phi_0'(\xi)\right)}{(t_s - t_\beta)\left(2\lambda + 2\phi_0(t_\beta) + (t_\beta - t_s)\phi_0'(t_\beta)\right)} \\ &= \frac{\mu - 1}{D(\Psi(\rho), \xi - t_s) - \mu} \cdot \frac{2\lambda + 2\phi_0(\xi) + (\xi - t_s)\phi_0'(\xi)}{2\lambda + 2\phi_0(t_\beta) + (t_\beta - t_s)\phi_0'(t_\beta)} \\ &\leq \frac{\mu - 1}{\rho_0 - \epsilon(\xi - t_s) - \mu} \cdot \frac{2\lambda + 2\phi_0(\xi) + \varphi_0(\xi)}{2\lambda - 2\phi_0(t_\beta^*) - \varphi_0(t_\beta^*)} \to \frac{\mu - 1}{\rho_0 - \mu} \text{ as } \xi \to t_s, \end{aligned}$$

where $t_\beta = t_\beta(\xi)$, $t_\beta^* = t_\beta^*(\xi)$, and $\varphi_0(\xi) = \sup\left\{\left|(\zeta - t_s)\phi_0'(\zeta)\right| : |\zeta - t_s| \leq |\xi - t_s|\right\}$. Combining this with (4.11) and (4.16), we can now choose $\delta_2 \in (0, \delta_4]$ such that $D_u\Psi(\rho)(\xi) \leq \tilde{K}$ for all $\rho \in \Omega(\delta_2)$, $\xi \in [t_s, t_s + \delta_2]$. □

From Lemmas 3.1, 4.1, and 4.2, we have the following theorem.

THEOREM 4.3. *$\Psi : \Omega(\delta) \to \Omega(\delta)$ for $\delta \leq \delta_2$.*

THEOREM 4.4. *There is some $\delta_* \in (0, \delta_2]$ such that Ψ is a contraction in the supremum norm on $\Omega(\delta_*)$.*

Proof. Suppose $\rho_1, \rho_2 \in \Omega(\delta_2)$, and let $t_{1,2}$, $\tilde{t}_{1,2}$ be defined as in the proof of Theorem 3.3. From (4.12), (4.13), and (4.14), we deduce that there is some $\delta_* \in (0, \delta_2]$ such that $t_{1,2} < \tilde{t}_{1,2}$ for any pair $\rho_1, \rho_2 \in \Omega(\delta_2)$. Thus $\tau < \tilde{\tau}$, which implies that $\phi_1'(\tilde{\tau}) < \phi_1'(\tau)$. Combining this with (3.8) gives

$$\|\Psi(\rho_1) - \Psi(\rho_2)\|_{\delta_*} \leq \frac{1}{\mu}\|\rho_1 - \rho_2\|_{\delta_*},$$

from which the theorem follows. ∎

From Theorems 4.3 and 4.4 and the completeness of $\Omega(\delta_*)$ in the supremum norm, we have the following theorem.

THEOREM 4.5. *$\Psi^n(\rho)$ converges to a unique solution of* (2.24) *in $\Omega(\delta_*)$ for any $\rho \in \Omega(\delta_*)$.*

Theorems 3.5 and 3.6 follow as before, except that, in the proof of Theorem 3.6, we can use 1 in place of κ.

5. Verification of loading/unloading near $t = t_s$. Throughout this section, we let $\rho(\xi)$ be the solution of (2.24) and $\sigma(x,t)$, $v(x,t)$ be continuous functions satisfying (1.2), (1.3), (2.1) on A_1 and A_0, and (2.2) on A_2. As mentioned in section 2, characteristic analysis can be used to find σ and v on A_1 and, once the interface $\zeta = \rho(\xi)$ between A_1 and A_2 is determined, on A_2. The goal of this section is to carry out such analysis and use it to verify that σ and v satisfy the original system (1.1) for at least a short time after the loading/unloading interface forms, i.e., that the solution is loading on A_1 and unloading on A_2. For simplicity, we assume in this section that ϕ is differentiable (except possibly at $t = t_s$). Then, by Theorem 3.6, ρ is differentiable, and it is not hard to show that σ and v are also then differentiable on A_1 and A_2. The following theorem shows that the solution is loading on A_1.

THEOREM 5.1. *$\partial_t \sigma(x,t) \le 0$ on $A_1 = \{(x,t) : s(t) < x < \beta t\}$.*

Proof. From (2.4), we have

$$(\sigma - \beta v)(x,t) = (\sigma - \beta v)(0, t - x/\beta)$$

on A_1. Combining this with (1.3) and (2.5), we have

$$\sigma(x,t) = -\phi(t - x/\beta) \tag{5.1}$$

on A_1. The theorem follows from (5.1) since ϕ is nondecreasing on $(0, t_\beta)$ and $0 < t - x/\beta < t_\beta$ on A_1. ∎

The following lemma gives a necessary and sufficient condition for the solution σ, v to be unloading in A_2. We should note that the condition can be checked only after the solution ρ of (2.24) is determined.

LEMMA 5.2. *$\partial_t \sigma(\xi,\zeta) \ge 0$ on $A_2 = \{(\xi,\zeta) : \xi < \zeta < \rho(\xi)\}$ iff*

$$\phi_1'(t_\beta(\xi))\left[\rho'(\xi) - \mu\right] \ge \phi_1'(\tilde{t}_\beta(\zeta))\left[\frac{1}{\mu} - \frac{1}{\rho'(\rho^{-1}(\zeta))}\right] \tag{5.2}$$

on A_2 or, equivalently,

$$\phi_1'(t_\beta(\xi))\left[\rho'(\xi) - \mu\right] \ge \phi_1'(t_\beta(\zeta))\left[\rho'(\zeta) - \mu\right] + \frac{(\mu-1)^2}{\mu}\phi_2'(\zeta) \tag{5.3}$$

on A_2. (See Figure 5.1.*)*

Proof. From (2.12), we have

$$\partial_\xi(\sigma + \alpha v) = 0,$$
$$\partial_\zeta(\sigma - \alpha v) = 0$$

on A_2. Combining this with (2.10), we have

$$(\sigma + \alpha v)(\xi,\zeta) = (\sigma + \alpha v)\left(\rho^{-1}(\zeta), \zeta\right) = \frac{2}{\mu - 1}\phi_1\left(\tilde{t}_\beta(\zeta)\right), \tag{5.4}$$
$$(\sigma - \alpha v)(\xi,\zeta) = (\sigma - \alpha v)(\xi, \rho(\xi)) = -\frac{2\mu}{\mu - 1}\phi_1(t_\beta(\xi)).$$

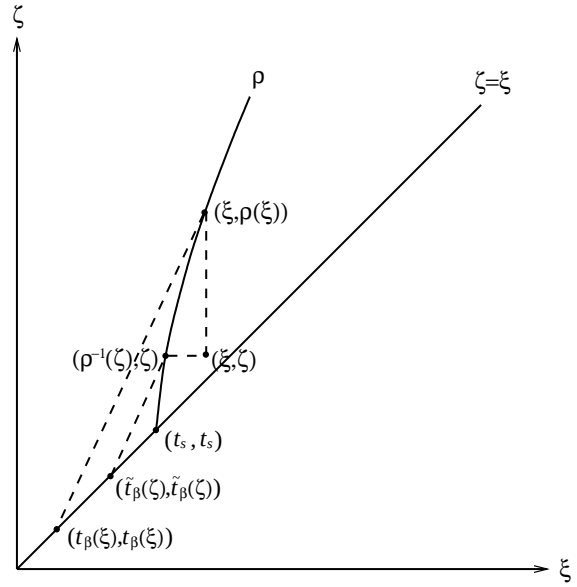


FIG. 5.1. *Derivation of the unloading condition.*

Adding these gives

$$\sigma(\xi,\zeta)=\frac{\phi_1\left(\tilde{t}_\beta(\zeta)\right)-\mu\phi_1\left(t_\beta(\xi)\right)}{\mu-1}. \tag{5.5}$$

Notice that $\partial_t=\partial_\xi+\partial_\zeta$, so

$$\begin{aligned}\partial_t\sigma(\xi,\zeta)&=\frac{\phi_1'\left(\tilde{t}_\beta(\zeta)\right)\tilde{t}_\beta'(\zeta)-\mu\phi_1'\left(t_\beta(\xi)\right)t_\beta'(\xi)}{\mu-1}\\&=\frac{\phi_1'\left(\tilde{t}_\beta(\zeta)\right)\left[\mu/\rho'\left(\rho^{-1}(\zeta)\right)-1\right]-\mu\phi_1'\left(t_\beta(\xi)\right)\left[\mu-\rho'(\xi)\right]}{(\mu-1)^2}\\&=\frac{\mu}{(\mu-1)^2}\left(\phi_1'\left(t_\beta(\xi)\right)\left[\rho'(\xi)-\mu\right]-\phi_1'\left(\tilde{t}_\beta(\zeta)\right)\left[\frac{1}{\mu}-\frac{1}{\rho'\left(\rho^{-1}(\zeta)\right)}\right]\right),\end{aligned}$$

from which (5.2) follows. From (3.14),

$$\phi_1'(\tilde{t}_\beta(\zeta))\left[\frac{1}{\mu}-\frac{1}{\rho'\left(\rho^{-1}(\zeta)\right)}\right]=\phi_1'(t_\beta(\zeta))\left[\rho'(\zeta)-\mu\right]+\frac{(\mu-1)^2}{\mu}\phi_2'(\zeta),$$

which implies (5.3). □

The following theorem shows that the solution is locally unloading on A_2.

THEOREM 5.3. *σ, v satisfy (1.1) on $\{(x,t):0<t<t_s+\delta\}$ for some $\delta>0$ in both the corner case and the smooth case.*

Proof. By Theorem 5.1 and Lemma 5.2, we need only show that (5.2) is satisfied on $\{(\xi,\zeta):\xi<\zeta<\rho(\xi),\ t_s<\xi<t_s+\delta\}$ for some $\delta>0$. Note that, in both the corner and smooth cases,

$$\rho'(t_s)-\mu\geq\frac{1}{\mu}-\frac{1}{\rho'(t_s)}.$$

The result then follows in the corner case from the fact that $\lim_{\xi \to t_s} \phi_1'(\xi) = \nu$ and in the smooth case from (4.16). □

6. Examples. In this section we apply the previously described iterative technique to (2.24) with $\mu = 3$ and specific boundary data; first,

$$\phi(t) = \begin{cases} \frac{1}{2} + \frac{1}{2}\sin\pi\left(t - \frac{1}{2}\right) & 0 \le t < 2, \\ 0 & t \ge 2; \end{cases} \tag{6.1}$$

then

$$\phi(t) = \frac{t}{t^2 + 1}. \tag{6.2}$$

(See Figures 6.1 and 6.2.) Notice that both fall under the smooth case and $t_s = 1$ for both. Our computations show that, with boundary data (6.1), $\partial_t \sigma(\xi, \zeta) < 0$ in part of A_2, and so σ, v do *not* satisfy (1.1) globally, while with boundary data (6.2), σ, v (as defined in section 5) satisfy (1.1) globally.

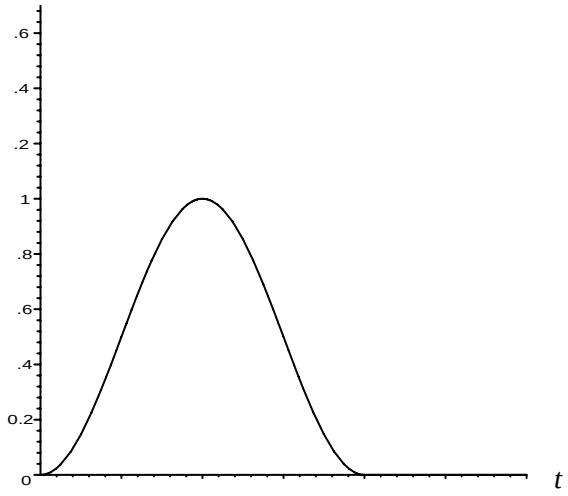


FIG. 6.1. *Boundary data* (6.1).

First, we use Maple V and the iterative technique described previously to compute the solution ρ of (2.24), (6.1). Figure 6.3 shows the graphs of $\mu\xi$ and ρ.

CLAIM 1. *Let σ, v be the solution of* (2.2) *with $\mu = 3$ satisfying* (1.3), (6.1), *and* (2.10). *Then $\partial_t \sigma(\xi, \zeta) < 0$ on a subset of $A_2 = \{(\xi, \zeta) : \xi < \zeta < \rho(\xi)\}$.*

Let

$$G(\xi) = \phi_1'(t_\beta(\xi))\left[\rho'(\xi) - \mu\right] \tag{6.3}$$

$$H(\xi) = \phi_1'(t_\beta(\xi))\left[\rho'(\xi) - \mu\right] + \frac{(\mu - 1)^2}{\mu}\phi_2'(\xi). \tag{6.4}$$

Graphs of G and H for (6.1) are shown in Figure 6.4. Notice that $G(\xi) = H(\xi)$ for $\xi \ge 2$ (this is because $\phi_2'(\xi) = 0$) and G is increasing for $\xi \ge 2$. This implies that $G(\xi) < H(\zeta) = G(\zeta)$ for $2 \le \xi < \zeta$, but, according to (5.3) and Lemma 5.2, $\partial_t \sigma(\xi, \zeta) \ge 0$ iff $G(\xi) \ge H(\zeta)$. This demonstrates the claim.

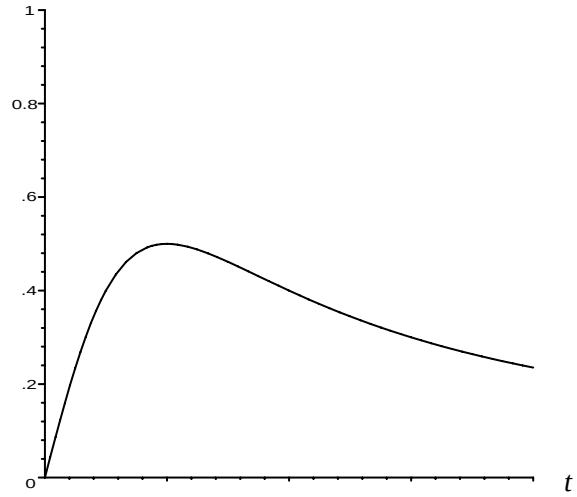


FIG. 6.2. *Boundary data* (6.2).

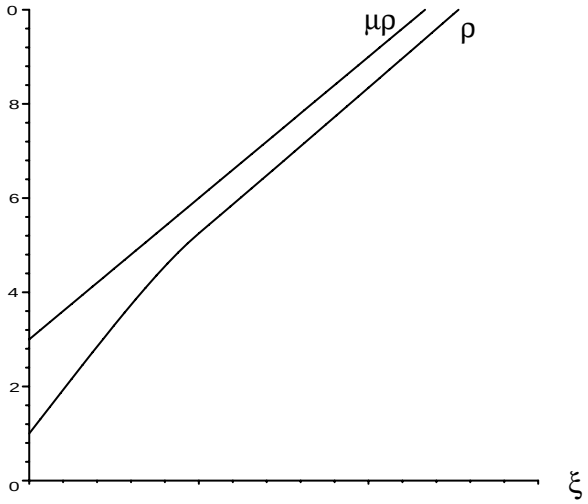


FIG. 6.3. *Solution of* (2.24), (6.1) *with* $\mu = 3$.

The claim is also illustrated by Figures 6.5 and 6.6. Figure 6.5 shows the graph of v (derived from (5.4)) at $t = 1.5$, 2, 2.5. The portion of each graph following the corner corresponds to the region A_2 and was derived assuming that the solution is unloading there. However, closer inspection of v at $t = 2.5$ (Figure 6.6) shows that $\partial_x v < 0$ (loading) on part of that region, thus invalidating the solution. Figure 6.7 shows the corresponding graphs of σ.

Next, we compute the solution ρ of (2.24), (6.2). Figure 6.8 shows the graphs of $\mu\xi$ and ρ.

CLAIM 2. *Let* σ, v *be the solution of* (2.2) *with* $\mu = 3$ *satisfying* (1.3), (6.2), *and* (2.10). *Then* $\partial_t \sigma(\xi, \zeta) \geq 0$ *on all of* $A_2 = \{(\xi, \zeta) : \xi < \zeta < \rho(\xi)\}$.

Graphs of G and H for (6.2) are shown in Figure 6.9. Notice that $G(\xi) > H(\xi)$

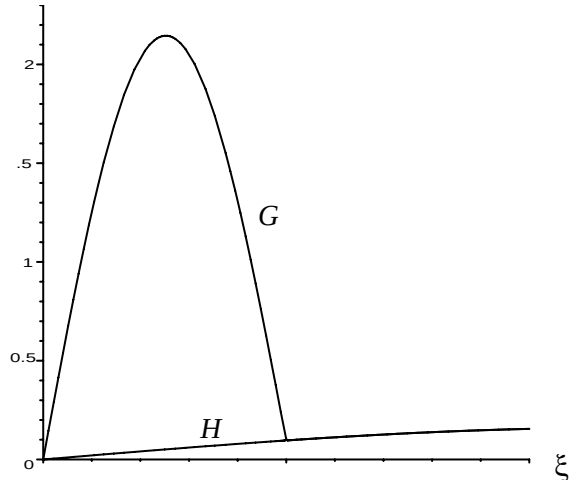


FIG. 6.4. *g and h with boundary data* (6.1).

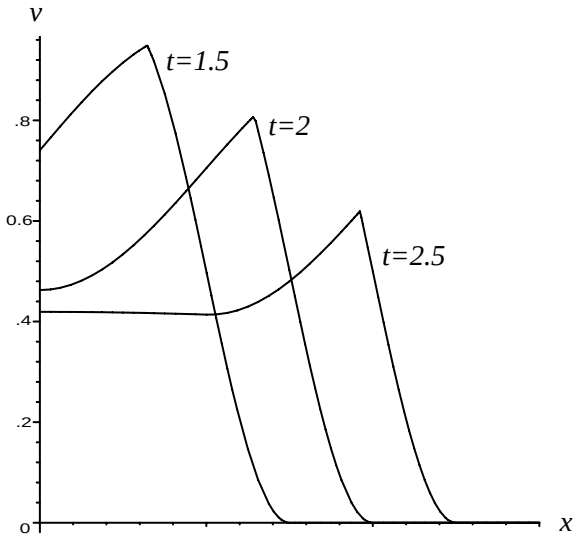


FIG. 6.5. *Solution of* (1.1), (1.2) *with boundary data* (6.1) *and* $\mu = 3$.

for $\xi \geq 1$ since $\phi_2'(\xi) < 0$ for $\xi \geq 1$. Let $H_{\max}(\xi) = \max\{H(\zeta) : \xi \leq \zeta \leq \rho(\xi)\}$. The claim is equivalent to $G(\xi) > H_{\max}(\xi)$ for $\xi > 1$. Figure 6.10 shows graphs of G and $H_{\max}$ for $1 \leq \xi \leq 10$, demonstrating the claim for ξ in that interval. Notice that $H = H_{\max}$ once H begins decreasing, so if we can show that H is decreasing for $\xi \geq 10$, the claim will follow. Toward that end, we will prove the following theorem.

THEOREM 6.1. *Suppose* ρ *is the solution of* (2.24), (6.2). *There is some* M *such that, if* $\xi_0 \geq M$ *and* H *is decreasing on* $[\xi_0, \rho(\xi_0)]$ *and*

$$H(\xi) \leq \frac{(\mu-1)^2 \ln \xi}{\mu \xi^2 \ln \mu} \tag{6.5}$$

on $[\xi_0, \rho(\xi_0)]$, *then* H *is decreasing on* $[\xi_0, \infty)$.

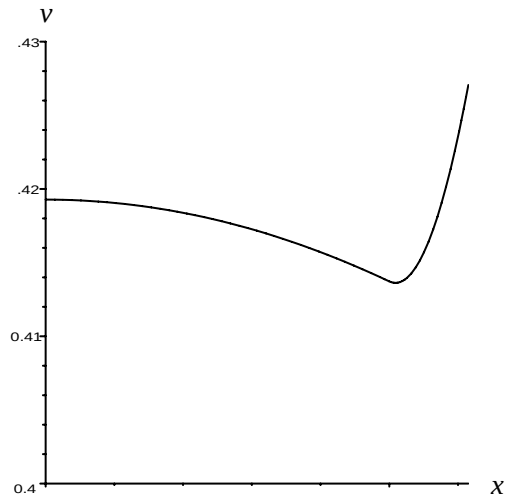


FIG. 6.6. $v(x, 2.5)$.

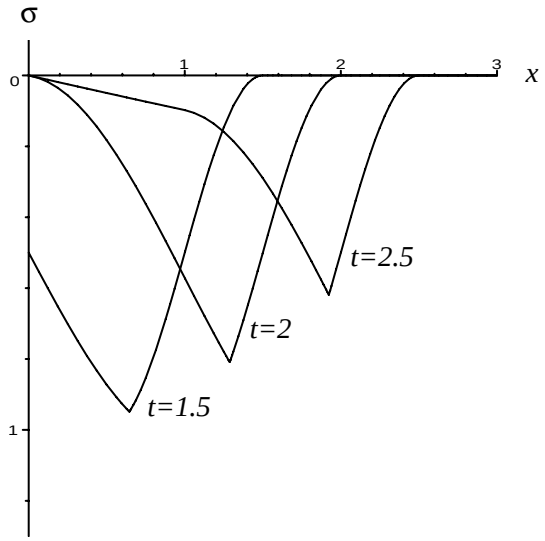


FIG. 6.7. *Solution* σ *of* (1.1), (1.2) *with boundary data* (6.1) *and* $\mu = 3$.

We begin with two lemmas.

LEMMA 6.2. *Suppose* ρ *is the solution of* (2.24), (6.2). *If* $\xi \geq \sqrt{e}$ *and* H *satisfies* (6.5) *on* $[\xi_0, \rho(\xi_0)]$, *then* H *satisfies* (6.5) *on* $[\xi_0, \infty)$.

Proof. Differentiating (2.24), we have

$$\mu\phi_1'(t_\beta)\left(\mu - \rho'(\xi)\right) - \phi_1'(\tilde{t}_\beta)\left(\frac{\mu - \rho'(\rho^{-1}(\xi))}{\rho'(\rho^{-1}(\xi))}\right) = (\mu - 1)^2\phi_2'(\xi).$$

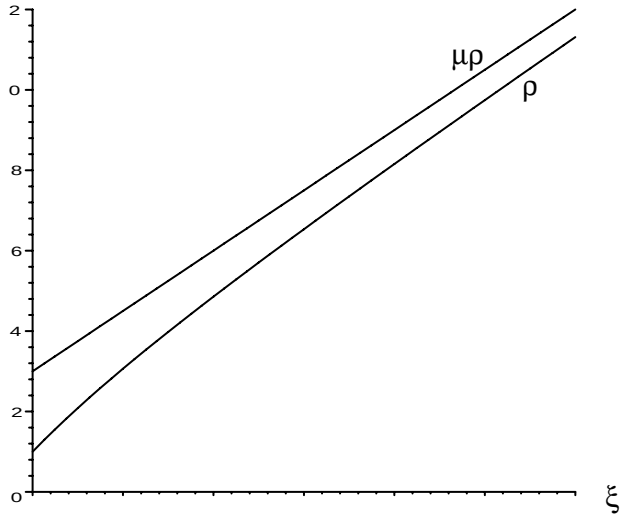


FIG. 6.8. *Solution of* (2.24), (6.2) *with* $\mu = 3$.

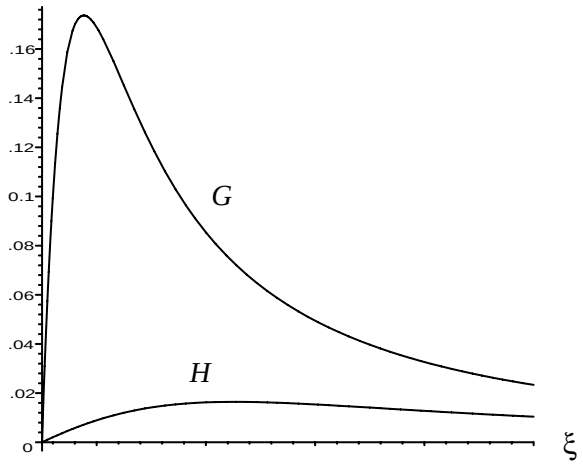


FIG. 6.9. *G and H with boundary data* (6.2).

Combining this with (6.4), we have

$$H(\xi) = \frac{\mu H(\rho^{-1}(\xi)) - (\mu-1)^2\phi_2'(\rho^{-1}(\xi))}{\mu^2\rho'(\rho^{-1}(\xi))},$$

which implies

$$H(\rho(\xi)) = \frac{\mu H(\xi) - (\mu-1)^2\phi_2'(\xi)}{\mu^2\rho'(\xi)} \leq \frac{\mu H(\xi) + (\mu-1)^2/\xi^2}{\mu^3} \tag{6.6}$$

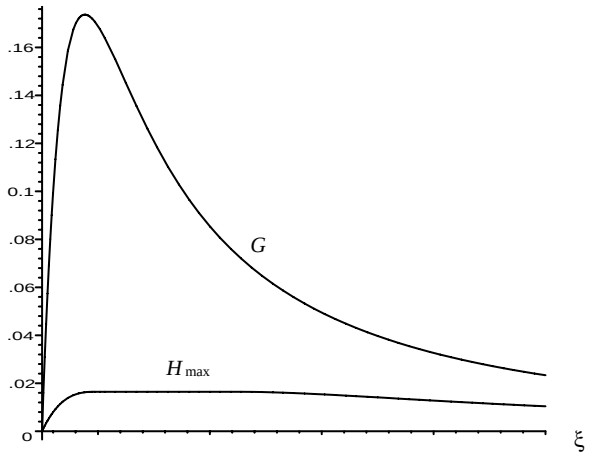


FIG. 6.10. *G and $H_{\max}$.*

by (2.25) and

$$-\phi_2'(\xi) = \frac{\xi^2 - 1}{\left(\xi^2 + 1\right)^2} < \frac{1}{\xi^2}. \tag{6.7}$$

Combining (6.5) and (6.6), we have

$$H(\rho(\xi)) \le \frac{(\mu-1)^2 \ln \xi}{\mu^3 \xi^2 \ln \mu} + \frac{(\mu-1)^2}{\mu^3 \xi^2} = \frac{(\mu-1)^2 \ln \mu\xi}{\mu(\mu\xi)^2 \ln \mu}.$$

Thus

$$H(\rho(\xi)) \le \frac{(\mu-1)^2 \ln \rho(\xi)}{\mu \rho(\xi)^2 \ln \mu}$$

for $\xi \ge \sqrt{e}/\mu$ by (2.26) and the fact that $(\ln \xi)/\xi^2$ is decreasing for $\xi \ge \sqrt{e}$. This means that H satisfies (6.5) on $\left[\rho(\xi_0), \rho^2(\xi_0)\right]$. Notice that (2.25) implies that $\lim_{n\to\infty} \rho^n(\xi_0) = \infty$ for $\xi_0 > 1$, so the result follows by induction. ∎

LEMMA 6.3. *Suppose ρ is the solution of* (2.24), (6.2) *satisfying* (6.5). *Then there is some M_1 such that*

$$\rho'(\xi) - \mu \le \frac{2(\mu-1)^2 \ln \mu\xi}{\mu \xi^2 \ln \mu}$$

for $\xi \ge M_1$.

Proof. From (6.4), we have

$$\rho'(\xi) - \mu = \frac{\mu H(\xi) - (\mu-1)^2 \phi_2'(\xi)}{\mu \phi_1'(t_\beta(\xi))}. \tag{6.8}$$

It is not hard to show that

$$\phi_1'(\xi) = \frac{1 - \xi^2}{\left(\xi^2 + 1\right)^2} \ge 1 - 3\xi^2 \ge \frac{1}{2} \quad \text{for } \xi \le \frac{1}{\sqrt{6}}. \tag{6.9}$$

Since t_β is decreasing by (2.25) and $\lim_{\xi\to\infty} t_\beta(\xi) = 0$ by (2.27), there is some M_1 such that

$$t_\beta(\xi) \le \frac{1}{\sqrt{6}} \quad \text{for } \xi \ge M_1. \tag{6.10}$$

Combining this with (6.7), (6.8), and (6.9) gives

$$\rho'(\xi) - \mu \le \frac{\mu H(\xi) + (\mu-1)^2/\xi^2}{\mu\left(1 - 3t_\beta(\xi)^2\right)} \le 2H(\xi) + \frac{2(\mu-1)^2}{\mu\xi^2}$$

for $\xi \ge M_1$. Combining this with (6.5) gives

$$\rho'(\xi) - \mu \le \frac{2(\mu-1)^2 \ln\xi}{\mu\xi^2 \ln\mu} + \frac{2(\mu-1)^2}{\mu\xi^2} = \frac{2(\mu-1)^2 \ln\mu\xi}{\mu\xi^2\ln\mu}$$

for $\xi \ge M_1$. □

We now give the proof of Theorem 6.1.

Proof. Assume that, on $[\xi_0, \rho(\xi_0)]$, H is decreasing and satisfies (6.5). Let $\xi_0 \le \xi < \zeta \le \rho(\xi_0)$. From (6.3), (6.4), and (6.6) we have

$$H(\rho(\xi)) = \frac{G(\xi)}{\mu\rho'(\xi)},$$

so

$$\begin{aligned} H(\rho(\xi)) - H(\rho(\zeta)) &= \frac{G(\xi)}{\mu\rho'(\xi)} - \frac{G(\zeta)}{\mu\rho'(\zeta)} = \frac{G(\xi)-G(\zeta)}{\mu\rho'(\xi)} + \frac{G(\zeta)}{\mu\rho'(\xi)\rho'(\zeta)}\left(\rho'(\zeta)-\rho'(\xi)\right) \\ &= \frac{G(\xi)-G(\zeta)}{\mu\rho'(\xi)} + \frac{G(\zeta)}{\mu\rho'(\xi)\rho'(\zeta)}\left(\frac{G(\zeta)}{\phi_1'(t_\beta(\zeta))} - \frac{G(\xi)}{\phi_1'(t_\beta(\xi))}\right) \quad \text{by (6.3)} \\ &= \frac{G(\xi)-G(\zeta)}{\mu\rho'(\xi)}\left[1 - \frac{G(\zeta)}{\rho'(\zeta)\phi_1'(t_\beta(\xi))}\right] + \frac{G(\zeta)^2\left(\phi_1'(t_\beta(\xi)) - \phi_1'(t_\beta(\zeta))\right)}{\mu\rho'(\xi)\rho'(\zeta)\phi_1'(t_\beta(\zeta))\phi_1'(t_\beta(\xi))}. \end{aligned}$$

The last term is negative since ϕ_1' and t_β are decreasing, so

$$H(\rho(\xi)) - H(\rho(\zeta)) \ge \frac{G(\xi)-G(\zeta)}{\mu\rho'(\xi)}\left(1 - 2G(\zeta)/\mu\right) - \frac{4G(\zeta)^2}{\mu^2}\left(\phi_1'(t_\beta(\zeta)) - \phi_1'(t_\beta(\xi))\right) \tag{6.11}$$

for $\xi \ge M_1$ by (2.25), (6.9), and (6.10). By (6.3), (6.4), (6.5), and (6.7), we have

$$G(\zeta) \le \frac{(\mu-1)^2\ln\zeta}{\mu\zeta^2\ln\mu} + \frac{(\mu-1)^2}{\mu\zeta^2} = \frac{(\mu-1)^2\ln\mu\zeta}{\mu\zeta^2\ln\mu}. \tag{6.12}$$

Combining this with Lemma 6.3, we can choose $M_2 \ge M_1$ such that

$$G(\zeta) \le \frac{\mu}{4} \text{ and } \rho'(\xi) \le \mu + 2\left(\frac{\mu}{4}\right) = \frac{3\mu}{2} \tag{6.13}$$

when $\xi \geq M_2$. Then, from (6.11), we have

$$\begin{aligned} H(\rho(\xi)) - H(\rho(\zeta)) &\geq \frac{G(\xi)-G(\zeta)}{3\mu^2} - \frac{4G(\zeta)^2}{\mu^2}\left(\phi_1'(t_\beta(\zeta)) - \phi_1'(t_\beta(\xi))\right) \\ &= \frac{H(\xi)-H(\zeta)}{3\mu^2} + \frac{(\mu-1)^2\left(\phi_2'(\zeta)-\phi_2'(\xi)\right)}{3\mu^3} \\ &\quad - \frac{4G(\zeta)^2}{\mu^2}\left(\phi_1'(t_\beta(\zeta)) - \phi_1'(t_\beta(\xi))\right) \\ &= \frac{H(\xi)-H(\zeta)}{3\mu^2} + \frac{(\mu-1)^2\phi_2''(c_1)(\zeta-\xi)}{3\mu^3} \\ &\quad - \frac{4G(\zeta)^2\phi_1''(c_2)t_\beta'(c_3)(\zeta-\xi)}{\mu^2}, \end{aligned} \tag{6.14}$$

where $\xi < c_1$, $c_3 < \zeta$, and $t_\beta(\zeta) < c_2 < t_\beta(\xi)$. It is not hard to show from (6.2) that

$$\phi_2''(\xi) \geq \frac{1}{\xi^3} \quad \text{for } \xi \geq \sqrt{12}, \tag{6.15}$$

$$\phi_1''(\xi) \geq -6\xi. \tag{6.16}$$

Combining (6.14), (6.15), and (6.16), we have

$$H(\rho(\xi)) - H(\rho(\zeta)) \geq \frac{H(\xi)-H(\zeta)}{3\mu^2} + \left[\frac{(\mu-1)^2}{3\mu^3 c_1^3} + \frac{24G(\zeta)^2 c_2 t_\beta'(c_3)}{\mu^2}\right](\zeta-\xi) \tag{6.17}$$

when $\xi \geq M_2, \sqrt{12}$. From (2.21) and Lemma 6.3, we have

$$t_\beta'(c_3) = \frac{\mu-\rho'(c_3)}{\mu-1} \geq -\frac{2(\mu-1)\ln \mu c_3}{\mu c_3^2 \ln \mu} \geq -\frac{2(\mu-1)\ln \mu\xi}{\mu\xi^2 \ln \mu}.$$

We now use this and (6.12) to estimate the quantity in brackets in (6.17):

$$\begin{aligned} \frac{(\mu-1)^2}{3\mu^3 c_1^3} + \frac{24G(\zeta)^2 c_2 t_\beta'(c_3)}{\mu^2} &\geq \frac{(\mu-1)^2}{3\mu^3 c_1^3} - \frac{48c_2(\mu-1)^5 \ln^2 \mu\zeta \ln \mu\xi}{\mu^5\zeta^4\xi^2 \ln^3 \mu} \\ &\geq \frac{(\mu-1)^2}{3\mu^3\zeta^3} - \frac{48t_\beta(\xi)(\mu-1)^5 \ln^2 \mu\zeta \ln \mu\xi}{\mu^5\zeta^4\xi^2 \ln^3 \mu} \\ &\geq \frac{(\mu-1)^2}{3\mu^3\zeta^3} - \frac{8\sqrt{6}(\mu-1)^5 \ln^2 \mu\zeta \ln \mu\xi}{\mu^5\zeta^4\xi^2 \ln^3 \mu} \end{aligned}$$

for $\xi \geq M_2, \sqrt{12}$ by (6.10). Combining this with (6.17), we have

$$\begin{aligned} H(\rho(\xi)) - H(\rho(\zeta)) &\geq \frac{H(\xi)-H(\zeta)}{3\mu^2} + \frac{(\mu-1)^2}{3\mu^3\zeta^3}\left[1 - \frac{24\sqrt{6}(\mu-1)^3 \ln^2 \mu\zeta \ln \mu\xi}{\mu^2\zeta\xi^2 \ln^3 \mu}\right](\zeta-\xi) \\ &\geq \frac{H(\xi)-H(\zeta)}{3\mu^2} + \frac{(\mu-1)^2}{3\mu^3\zeta^3}\left[1 - \frac{24\sqrt{6}(\mu-1)^3 \ln^3 \mu\xi}{\mu^2\xi^3 \ln^3 \mu}\right](\zeta-\xi) \end{aligned} \tag{6.18}$$

for $\xi \geq e^2/\mu$ since $(\ln^2 \mu\zeta)/\zeta$ is then decreasing. Since the function in brackets is increasing for $\xi \geq e/\mu$ and approaches 1 as $\xi \to \infty$, we can choose $M \geq M_2, \sqrt{12}, e^2/\mu$

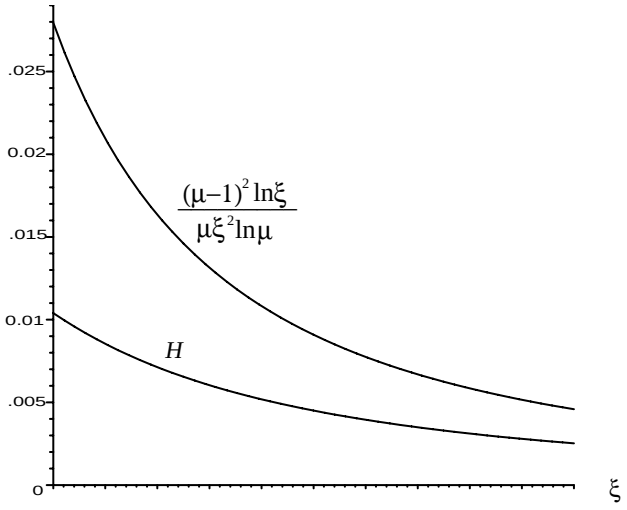


FIG. 6.11. *Bound on H.*

so that

$$1 - \frac{24\sqrt{6}(\mu-1)^2 \ln^3 \mu\xi}{\mu^2 \xi^3 \ln^3 \mu} > 0 \tag{6.19}$$

for $\xi \geq M$. Since H is decreasing on $[\xi_0, \rho(\xi_0)]$, (6.18) implies that H is decreasing on $[\rho(\xi_0), \rho^2(\xi_0)]$ if $\xi_0 \geq M$. The result follows by induction and Lemma 6.2. ∎

Our computations with $\mu = 3$ show that (6.10), (6.13), and (6.19) all hold for $\xi \geq 10 > \sqrt{12}, e^2/\mu$, so we can let $M = 10$. Figure 6.11 shows that (6.5) holds on $[10, \rho(10)]$ and that H is decreasing on $[10, \rho(10)]$, and so Theorem 6.1 implies that H is decreasing on $[10, \infty)$, from which Claim 2 follows.

Figures 6.12 and 6.13 show the graphs of v and σ at $t = 2, 5, 10$. The solution of (1.1), (1.2), (1.3), (6.2) with $\mu = 3$ is qualitatively the same as for the stress-controlled problem in [3], i.e., a decaying pulse of increasing length consisting of a loading front followed by an unloading front. We prove the decay in the following theorem.

THEOREM 6.4. *Let σ, v be the solution of* (2.2) *with $\mu = 3$ satisfying* (1.3), (6.2)*, and* (2.10). *Then*

$$\lim_{t\to\infty} \max_x \{|v(x,t)|, |\sigma(x,t)|\} = 0.$$

Proof. Since $\partial_x v$ changes sign only at the loading/unloading interface, we have

$$\max_x |v(x,t)| = v(s(t), t) = \phi(t - s(t)/\beta)/\beta \tag{6.20}$$

by (2.10). Equation (2.27) implies that $\lim_{t\to\infty}(t - s(t)/\beta) = 0$, and hence

$$\lim_{t\to\infty} \max_x |v(x,t)| = 0.$$

Now, from (5.1), we have

$$\partial_x \sigma(x,t) = \phi'(t - x/\beta)/\beta$$

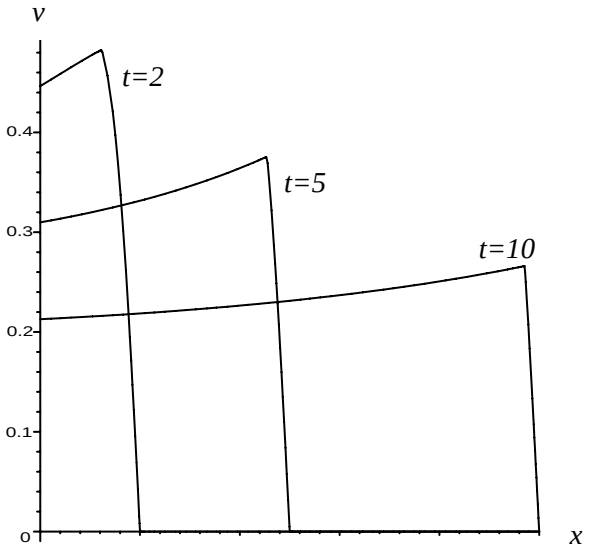


FIG. 6.12. *Solution v of* (1.1), (1.2) *with boundary data* (6.2) *and* $\mu = 3$.

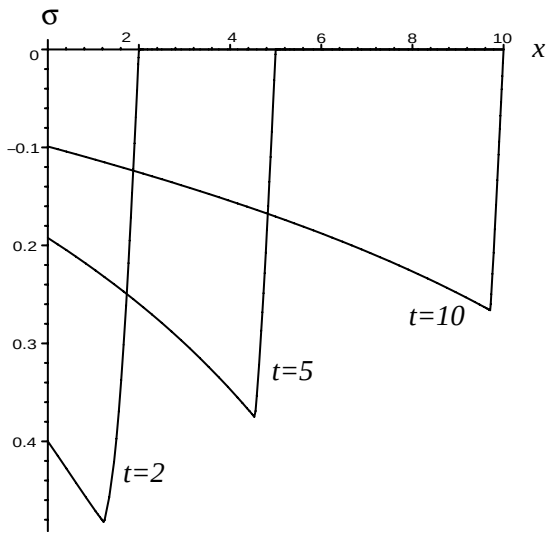


FIG. 6.13. *Solution σ of* (1.1), (1.2) *with boundary data* (6.2) *and* $\mu = 3$.

which is nonnegative on A_1 since

$$t - x/\beta \le t - s(t)/\beta = t_\beta \le t_s$$

on A_1. This implies that

$$0 \ge \sigma(x,t) \ge \sigma(s(t),t) = -\phi(t - s(t)/\beta)$$

by (2.10), and so

$$\lim_{t\to\infty} \max_{x\ge s(t)} |\sigma(x,t)| = 0.$$

We now focus on A_2. Combining (5.5) and (2.24), we have

$$\sigma(\xi,\zeta) = \frac{\mu}{\mu-1}\left[\phi_1\left(t_\beta(\zeta)\right) - \phi_1\left(t_\beta(\xi)\right)\right] - \phi_2(\zeta), \tag{6.21}$$

which is negative on A_2 since t_β is nonincreasing (by (2.21) and (2.25)), ϕ_1 is nondecreasing, and $\zeta > \xi$. Also, (6.21) implies that

$$\begin{aligned}\sigma(\xi,\zeta) &\geq -\frac{\mu}{\mu-1}\phi_1\left(t_\beta(\xi)\right) - \phi_2(\zeta)\\ &= -\frac{\mu}{\mu-1}\phi_1\left(t_\beta(t-x/\alpha)\right) - \phi_2(t+x/\alpha)\\ &\geq -\frac{\mu}{\mu-1}\phi_1\left(t_\beta(t-s(t)/\alpha)\right) - \phi_2(t)\end{aligned}$$

on A_2 since ϕ_2 and t_β are nonincreasing and ϕ_1 is nondecreasing. By (2.7),

$$t_\beta(t-s(t)/\alpha) = t - s(t)/\beta,$$

so

$$0 \geq \sigma(x,t) \geq -\frac{\mu}{\mu-1}\phi_1\left(t-s(t)/\beta\right) - \phi_2(t).$$

The right-hand side goes to zero as $t \to \infty$ by (2.27), so

$$\lim_{t\to\infty}\max_{0\leq x\leq s(t)}|\sigma(x,t)| = 0. \qquad \square$$

Acknowledgments. We thank David Schaeffer and Michael Shearer for their many helpful comments made on an earlier draft of this paper.

REFERENCES

[1] E. Bauer, *Calibration of a comprehensive hypoplastic model for granular materials, soils and foundations,* Jap. Soc. Soil Mech. Found. Eng., 36 (1996), pp. 13–26.

[2] M. K. Gordon, *Perturbed scale-invariant initial value problems in one-dimensional dynamic elastoplasticity,* SIAM J. Math. Anal., 26 (1995), pp. 1564–1587.

[3] M. S. Gordon, M. Shearer, and D. Schaeffer, *Plane shear waves in a fully saturated granular material with velocity and stress controlled boundary conditions,* Internat. J. Non-linear Mech., 32 (1997), pp. 489–503.

[4] G. Gudehus, *A comprehensive constitutive equation for granular materials,* Soils and Foundations, 36 (1996), pp. 1–12.

[5] D. Kolymbas and W. Wu, *Introduction to Hypoplasticity,* in Modern Approaches to Plasticity, D. Kolymbas, ed., Elsevier, Amsterdam, 1993, pp. 213–223.

[6] D. Kolymbas, I. Herle, and P. A. von Wolffersdorff, *Hypoplastic constitutive equation with internal variables,* Internat. J. Numer. Anal. Meth. Geomech., 19 (1995), pp. 415–436.

[7] V. A. Osinov and G. Gudehus, *Plane shear waves and loss of stability in a saturated granular body,* Mechanics of Cohesive Frictional Materials and Structures, 1 (1996), pp. 25–44.

SIAM J. MATH. ANAL.
Vol. 34, No. 3, pp. 555–572

ASYMPTOTIC STABILITY OF TRAVELING WAVES TO A CERTAIN DISCRETE VELOCITY MODEL OF THE BOLTZMANN EQUATION IN THE HALF SPACE*

SHINYA NISHIBATA†

Dedicated to Professors Takaaki Nishida and Masayasu Mimura on their 60th birthdays

Abstract. The present paper studies the asymptotic stability of a traveling wave for the Broadwell model in a half space. This model admits the traveling wave which connects two distinct Maxwellian states at the spatial asymptotic points. The traveling wave is shown to be time asymptotically stable if the fluid dynamical velocity is less than a certain positive value. This stability theorem is proved by applying the standard energy method. Here, the location of the traveling wave, which should be a time asymptotic state, is shifted by boundary effect. This shift is estimated by utilizing the property that the traveling wave converges to the Maxwellian states exponentially fast.

Key words. Broadwell model, boundary effect, initial boundary value problem, energy method

AMS subject classifications. 35B35, 35B40, 76N15

PII. S0036141001390683

1. Introduction.

1.1. Problems. We study the asymptotic stability of a traveling wave solution to the Broadwell model system

$$\partial_t F_1 + v\partial_x F_1 = q(F), \tag{1.1a}$$

$$4\partial_t F_2 = -2q(F), \tag{1.1b}$$

$$\partial_t F_3 - v\partial_x F_3 = q(F) \tag{1.1c}$$

in the first half space $\mathbb{R}_+ := \{x > 0\}$. Here,

$$q(F) := F_2^2 - F_1 F_3, \quad F := (F_1, F_2, F_3),$$

v is a positive constant, and unknown functions $F_i > 0$ for $i = 1, 2, 3$ represent the mass densities for gas particles moving with the speed v, 0, and $-v$ in the x-direction, respectively.

We prescribe the initial condition

$$F(x,0) = F_0(x) = (F_{1,0}, F_{2,0}, F_{3,0})(x), \tag{1.2a}$$

$$F_0(x) \to M^+ = (M_1^+, M_2^+, M_3^+) \quad \text{as} \quad x \to \infty, \tag{1.2b}$$

where M^+ is a Maxwellian state. The Maxwellian state $M := (M_1, M_2, M_3)$ is an equilibrium state of (1.1) with positive entries

$$q(M) = 0, \quad M_i > 0 \text{ for } i = 1,\ 2,\ 3. \tag{1.3}$$

*Received by the editors June 8, 2001; accepted for publication (in revised form) May 13, 2002; published electronically December 13, 2002. This work was supported in part by Grant-in-Aid for Scientific Research (A) 12740116 of the Ministry of Education, Science, Sports and Culture.
http://www.siam.org/journals/sima/34-3/39068.html
†Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo 152-8552, Japan (shinya@is.titech.ac.jp).

Since the characteristic speed v of F_1 is positive, it is necessary to set one condition on the boundary $\{x = 0\}$ for the well-posedness of the initial and boundary value problem (1.1) and (1.2). Then we adopt the pure diffusion boundary condition as

$$F_1(0,t) = B_1, \quad B_1 > 0, \tag{1.4}$$

where B_1 is a constant. The compatibility condition of order zero is supposed to hold:

$$F_{1,0}(0) = B_1. \tag{1.5}$$

The traveling wave to the system (1.1) is a solution in the form of $\tilde{F}(\xi) := (\tilde{F}_1, F_2, F_3)(\xi)$ interpolating two distinct Maxwellian states $M^{\pm} = (M_1^{\pm}, M_2^{\pm}, M_3^{\pm})$:

$$\tilde{F}(\xi) \to M^{\pm} = (M_1^{\pm}, M_2^{\pm}, M_3^{\pm}) \text{ as } \xi \to \pm\infty, \quad M^+ \neq M^-, \tag{1.6}$$

where $\xi := x - st$ and s is a constant called the traveling wave speed. The main purpose of the present paper is to show the asymptotic stability of the traveling wave with $s > 0$ in the first half space $\mathbb{R}_+$. It is shown in Lemma 1.1 that for a positive constant B_1 and a Maxwellian state M^+, if $B_1 > M_1^+$, then we can find a unique Maxwellian state $M^- = (M_1^-, M_2^-, M_3^-)$ such that there exists a traveling wave $\tilde{F}(x - st)$, with $s > 0$, which satisfies (1.6) with $M_1^- = B_1$.

1.2. Equations for traveling waves. Substituting $\tilde{F}(\xi)$ in (1.1) yields that

$$(v - s)\tilde{F}_1' = \tilde{F}_2^2 - \tilde{F}_1\tilde{F}_3, \tag{1.7a}$$

$$-4s\tilde{F}_2' = -2(\tilde{F}_2^2 - \tilde{F}_1\tilde{F}_3), \tag{1.7b}$$

$$-(v + s)\tilde{F}_3' = \tilde{F}_2^2 - \tilde{F}_1\tilde{F}_3. \tag{1.7c}$$

Here and hereafter, the superscript " $'$ " denotes the differentiation with respect to ξ. From (1.7), it holds that

$$\{(v - s)\tilde{F}_1 - 4s\tilde{F}_2 - (v + s)\tilde{F}_3\}' = 0, \tag{1.8a}$$

$$\{(v - s)\tilde{F}_1 + (v + s)\tilde{F}_3\}' = 0. \tag{1.8b}$$

Suppose that the traveling wave solution $\tilde{F}(\xi)$, $\xi \in \mathbb{R}$, satisfying (1.6) with (1.3), exists for the moment. Integrating (1.8) over $(-\infty, \infty)$ and $(-\infty, \xi]$, respectively, we have

$$(v - s)\tilde{F}_1 - 4s\tilde{F}_2 - (v + s)\tilde{F}_3 = r_1, \quad r_1 := (v - s)M_1^{\pm} - 4sM_2^{\pm} - (v + s)M_3^{\pm}, \tag{1.9a}$$

$$(v - s)\tilde{F}_1 + (v + s)\tilde{F}_3 = r_2, \quad r_2 := (v - s)M_1^{\pm} + (v + s)M_3^{\pm}. \tag{1.9b}$$

The right equalities in (1.9) are the Rankine–Hugoniot condition. It is easy to see from (1.7) that $s \neq \pm v$, 0. Represent $\tilde{F}_2$ and $\tilde{F}_3$ in terms of $\tilde{F}_1$ by solving (1.9) and then substitute the resultant expressions in (1.7a). Apply the same procedure to $\tilde{F}_2$ and $\tilde{F}_3$, too. The results are

$$\tilde{F}_1' = \frac{v^2 + 3s^2}{4s^2(v + s)}(\tilde{F}_1 - M_1^-)(\tilde{F}_1 - M_1^+), \tag{1.10a}$$

$$\tilde{F}_2' = \frac{v^2 + 3s^2}{4s(v + s)(v - s)}(\tilde{F}_2 - M_2^-)(\tilde{F}_2 - M_2^+), \tag{1.10b}$$

$$\tilde{F}_3' = -\frac{v^2 + 3s^2}{4s^2(v - s)}(\tilde{F}_3 - M_3^-)(\tilde{F}_3 - M_3^+). \tag{1.10c}$$

Since $\tilde{F}(\xi)$ satisfies (1.6), we arrive at the inequality

$$0 < s < v \tag{1.11a}$$

$$\text{or} \quad -v < s < 0. \tag{1.11b}$$

It is easy to see from (1.10) that the traveling wave is monotonic. Especially, $\tilde{F}_1(\xi)$ is monotonically decreasing and thus $M_1^+ < M_1^-$. $\tilde{F}_2(\xi)$ is monotonically decreasing if and only if (1.11a) holds.

The above observation means that the condition (1.11) is necessary for the existence of the traveling wave $\tilde{F}(\xi)$. Moreover, it is proved in [2], the condition (1.11) is sufficient for the existence of the traveling wave. This fact is also proved by the straightforward computation with (1.10). In the next lemma, we summarize the results concerning the existence of the traveling wave in a convenient formulation to the present paper. The proof follows from algebraic computations with (1.3), (1.9), and (1.11).

LEMMA 1.1. (i) *If there exists a traveling wave* $\tilde{F}(\xi)$, $\xi := x - st \in \mathbb{R}$, *interpolating two distinct Maxwellian states* $M^{\pm}$, *then* (1.11) *hold. Moreover, if* $s \lesseqgtr 0$, *then* $M_2^- \lesseqgtr M_2^+$.

(ii) *Suppose that a positive constant* B_1 *and a Maxwellian state* $M^+ = (M_1^+, M_2^+, M_3^+)$ *satisfy* $B_1 > M_1^+$. *If* $|B_1 - M_1^+|$ *is sufficiently small, then there exists a unique Maxwellian state* $M^- = (M_1^-, M_2^-, M_3^-)$ *such that* $M_1^- = B_1$, $M_2^- > M_2^+$, *and the Rankine–Hugoniot condition in* (1.9) *holds. Therefore, there exists a traveling wave* $\tilde{F}(x - st)$, *uniquely up to a shift, which interpolates* $M^{\pm}$ *and satisfies* (1.11a).

Due to Lemma 1.1, we assume that

$$M_1^+ < B_1 \tag{1.12}$$

and (1.11a) hold, here and hereafter. For a given Maxwellian state M^+ and B_1 satisfying (1.12), M^- is determined by the algebraic relation (1.9) and (1.3) with $M_1^- = B_1$. Therefore, it holds that

$$|M_i^+ - M_i^-| \le C\delta_M, \quad \delta_M := |M_1^+ - B_1| \quad \text{for } i = 1, 2, 3. \tag{1.13}$$

We often call the quantity δ_M the shock strength.

1.3. Fluid dynamical equations. We rewrite the system (1.1) in terms of the fluid dynamical quantities, following the context of [3]. The density ρ and the momentum m are defined by

$$\rho := F_1 + 4F_2 + F_3, \quad m := v(F_1 - F_3). \tag{1.14a}$$

In addition, we denote

$$z := v^2(F_1 + F_3). \tag{1.14b}$$

From (1.1) and (1.14), we have the system of equations for the fluid dynamical quantities

$$\rho_t + m_x = 0, \tag{1.15a}$$

$$m_t + z_x = 0, \tag{1.15b}$$

$$z_t + v^2 m_x = \frac{(v^2\rho - z)^2 - 4(z^2 - v^2 m^2)}{8v^2}. \tag{1.15c}$$

The initial data for the above system (1.15) is derived from (1.14) and written as

$$\begin{aligned} \rho(x,0) &= \rho_0(x) := (F_{1,0} + 4F_{2,0} + F_{3,0})(x), \\ m(x,0) &= m_0(x) := v(F_{1,0} - F_{3,0})(x), \\ z(x,0) &= z_0(x) := v^2(F_{1,0} + F_{3,0})(x). \end{aligned} \tag{1.16}$$

The spatial asymptotic states for the fluid dynamical quantities are given by

$$\rho^\pm = M_1^\pm + 4M_2^\pm + M_3^\pm > 0, \quad m^\pm := v(M_1^\pm - M_3^\pm), \quad z^\pm := v^2(M_1^\pm + M_3^\pm) > 0. \tag{1.17}$$

The above inequalities follow from the positivity of each $M_i^\pm$. The condition (1.3) for the Maxwellian state M is rewritten in the fluid dynamical quantities as

$$|u| < v, \quad z = \rho\sigma(u), \quad \sigma(u) := \frac{v^2}{3}\left(2\sqrt{1 + \frac{3u^2}{v^2}} - 1\right), \tag{1.18}$$

where $u := m/\rho$ is called the fluid dynamical velocity. Since the asymptotic states $(\rho_\pm, m_\pm, z_\pm)$ are Maxwellian states, $(\rho_\pm, m_\pm, z_\pm)$ satisfy (1.18).

We express the traveling wave solution to (1.15) by

$$(\tilde\rho, \tilde m, \tilde z) := (\tilde F_1 + 4\tilde F_2 + \tilde F_3, v(\tilde F_1 - \tilde F_3), v^2(\tilde F_1 + \tilde F_3)), \tag{1.19}$$

which satisfies

$$-s\tilde\rho' + \tilde m' = 0, \tag{1.20a}$$

$$-s\tilde m' + \tilde z' = 0, \tag{1.20b}$$

$$-s\tilde z' + v^2\tilde m' = \frac{(v^2\tilde\rho - \tilde z)^2 - 4(\tilde z^2 - v^2\tilde m^2)}{8v^2}. \tag{1.20c}$$

The existence of the traveling wave $(\tilde\rho,\ \tilde m,\ \tilde z)$ immediately follows from (1.19) and the existence of the traveling wave $(\tilde F_1, \tilde F_2, \tilde F_3)$ to (1.7) in Lemma 1.1. Substituting (1.19) in (1.9), we have the Rankine–Hugoniot condition

$$-s(\rho^+ - \rho^-) + (m^+ - m^-) = 0, \tag{1.21a}$$

$$-s(m^+ - m^-) + (z^+ - z^-) = 0. \tag{1.21b}$$

Apparently, from (1.13),

$$|(\rho^+ - \rho^-, m^+ - m^-, z^+ - z^-)| \le C\delta_M. \tag{1.22}$$

It is shown in [3] that the condition (1.11) is equivalent to the Lax entropy condition

$$\lambda_-(u^+) < s < \lambda_+(u^-) \quad \text{or} \quad \lambda_+(u^+) < s < \lambda_-(u^-), \tag{1.23}$$

where

$$\lambda_\pm(u) := \frac{u \pm \sqrt{\sigma(u)}}{\sqrt{1 + 3(u^2/v^2)}}, \quad u^\pm := \frac{m^\pm}{\rho^\pm}. \tag{1.24}$$

Here, $\lambda_\pm(u)$ are the characteristics of the corresponding Euler equation

$$\rho_t + m_x = 0, \tag{1.25a}$$

$$m_t + (\rho\sigma(u))_x = 0. \tag{1.25b}$$

The system (1.25) is derived from (1.15) through the Chapman–Enskog expansion.

1.4. Assumptions and the main result. To handle the boundary terms, the given Maxwellian state M^+ is supposed to satisfy

$$M_1^+ - M_3^+ < 2M_2^+. \tag{1.26}$$

The condition (1.26) apparently holds if the fluid dynamical momentum is negative at the Maxwellian state M^+. The condition (1.26) with M^+ replaced by $\tilde{F}(\xi)$ also holds, provided that the traveling wave $\tilde{F}(\xi)$ interpolates M^- and M^+ with sufficiently small shock strength, $\delta_M \ll 1$. The condition (1.26) is rewritten in the fluid dynamical quantities as

$$u^+ < \frac{2-\sqrt{2}}{2} v, \tag{1.26$'$}$$

where we have used (1.18).

Since the traveling wave $\tilde{F}(\xi)$ converges to M^+ exponentially fast as ξ tends to infinity, we can define the antiderivatives of the initial perturbations from the traveling wave if $F_0 - M_+ \in L^1(\mathbb{R}_+)$. Thus, we define

$$\hat{\Phi}_0(x) := -\int_x^\infty \hat{\phi}_0(x)dy, \quad \hat{\Psi}_0(x) := -\int_x^\infty \hat{\psi}_0(x)dy, \tag{1.27a}$$

$$\begin{aligned} &\hat{\phi}_0(x) := \rho_0(x) - \tilde{\rho}(x-\beta), \quad \hat{\psi}_0(x) := m_0(x) - \tilde{m}(x-\beta), \\ &\hat{\omega}_0(x) := z_0(x) - \tilde{z}(x-\beta), \end{aligned} \tag{1.27b}$$

where β is a positive constant. The parameter β is determined later and utilized to handle the boundary terms in subsection 2.3.

THEOREM 1.2. *Suppose that* (1.5) *and* (1.26) *hold. Let* $F_0 - M^+ \in (L^1 \cap H^1)(\mathbb{R}_+)$ *and* $(\hat{\Phi}_0, \hat{\Psi}_0) \in L^2(\mathbb{R}_+)$. *Then there exists a constant* $\bar{\delta}_0$ *with the following property: if* $\delta_M \le \bar{\delta}_0$, *then one can find a positive constant* β_0 *such that whenever* $\beta \ge \beta_0$ *and* $\|(\hat{\Phi}_0, \hat{\Psi}_0)\| + \|F_0(\cdot) - \tilde{F}(\cdot - \beta)\|_1$ *is sufficiently small, the initial boundary value problem* (1.1), (1.2), *and* (1.4) *has a unique solution* $F(x,t)$ *globally in time. Moreover, the solution* $F(x,t)$ *satisfies that for a certain constant* x_∞,

$$\sup_{x \in \mathbb{R}_+} |F(x,t) - \tilde{F}(x - st + x_\infty)| \to 0 \quad as \quad t \to \infty. \tag{1.28}$$

The Broadwell model (1.1) is a simple but typical model of the Boltzmann equation. The existence of the traveling wave to the model is established by Caflish [2]. It is proved by Kawashima and Matsumura in [3] that the traveling wave is asymptotically stable in the full space $\mathbb{R}$. In [3], the energy method is employed to obtain the a priori estimate. In the present research, we also use this method and follow some computations of their research. Due to Lemma 1.1, we have to assume $B_1 > M_1^+$ for the existence of the traveling wave. The opposite case $B_1 < M_1^+$ gives the solution which converges to the rarefaction wave in the half space $\mathbb{R}_+$. It has been recently proved by Nikkuni and Kawashima in [9].

The half space problem for general systems of the Boltzmann equation with discrete velocities is considered in [4], [5], [8], and [10]. These papers are concerned with the stationary wave, which is the traveling wave with zero speed. The research in [10] and [4] shows the existence of the stationary wave with the pure diffusive boundary condition. The stability of this stationary wave is proved in [8]. The existence and the stability of the stationary wave with the reflective boundary condition are proved in [4].

Following these results, the present research studies the stability of the traveling wave with positive speed, $s > 0$. Although these two types of waves, the traveling wave and the stationary wave, look similar, the positivity of the traveling wave speed gives rise to some essential difficulties. For example, it has been enough to consider a certain part of the stationary wave in [4] and [8]. As the traveling wave moves in a positive direction, the present research needs to handle the whole shape of the wave. Thus, we have to introduce the antiderivatives of perturbations from the traveling wave. Furthermore, we cannot determine the asymptotic location of the traveling wave by the initial condition. Namely, the boundary effect causes a shift of the location. This difficulty is resolved by employing the idea in Matsumura and Mei [7], where the same kind of problem is considered for the viscous p-system.

The plan of the present paper is as follows. In the next section, we discuss the property of the traveling wave and then derive the system of equations to the antiderivatives of perturbations in terms of fluid dynamical quantities. Then we show the time exponential decay of a certain linear combination of the boundary data by utilizing the decay property of the traveling wave at the spatial asymptotic points. In section 3, we obtain the a priori estimates by the energy method applied to the fluid dynamical equations. In these computations, the decay property of the boundary condition plays an essential role. Here, we are indebted to some ideas in the preceding research in [3] and [7].

Notation. For $1 \le p \le \infty$, $L^p(\Omega)$ denotes the usual Lebesgue space over Ω with the norm $|\cdot|_p$. For an arbitrary integer $l \ge 0$, H^l denotes the lth order Sobolev space in the L^2-sense, equipped with the norm $\|\cdot\|_l$. We note $H^0 = L^2$ and $\|\cdot\| := \|\cdot\|_0 = |\cdot|_2$. We also denote by $C^k(I; H^l(\Omega))$ the space of k-times continuously differentiable functions on the interval I with values in $H^l(\Omega)$. Finally, by c and C we denote several constants without confusion.

2. Preliminary calculation.

2.1. Properties of traveling waves. In this subsection, we summarize the property of the traveling wave satisfying (1.11a). We first normalize the traveling wave as

$$\tilde{F}_1(0) = \frac{1}{2}(M_1^+ + M_1^-) \tag{2.1}$$

for clarity. Then it holds from (1.10a) that, for $\xi \in \mathbb{R}$,

$$\begin{cases} 0 < M_1^- - \tilde{F}_1(\xi) \le (M_1^- - M_1^+)e^{\sigma\xi}/2 & \text{for} \quad \xi \ge 0, \\ 0 < \tilde{F}_1(\xi) - M_1^+ \le (M_1^- - M_1^+)e^{-\sigma\xi}/2 & \text{for} \quad \xi < 0, \end{cases} \tag{2.2a}$$

$$\sigma := \frac{v^2 + 3s^2}{4s^2(v+s)}(M_1^- - M_1^+) > 0, \tag{2.2b}$$

where we have used $s > 0$. From (1.9) and (1.14), we have

$$-s\tilde{\rho} + \tilde{m} = -s\rho_\pm + m_\pm, \quad -s\tilde{m} + \tilde{z} = -sm_\pm + z_\pm. \tag{2.3}$$

Substituting (2.3) in (1.20) yields that

$$\tilde{\rho}' = C_\rho(\tilde{\rho} - \rho_+)(\tilde{\rho} - \rho_-), \quad C_\rho := \frac{v^2 + 3s^2}{8sv^2} > 0, \tag{2.4}$$

where we have used (1.11a). It holds from (1.22), (2.2a), and (2.4) that

$$\tilde{\rho}' < 0, \quad \tilde{m}' = s\tilde{\rho}' < 0, \quad \tilde{z}' = s^2\tilde{\rho}' < 0, \tag{2.5a}$$

$$|\tilde{\rho}(\xi) - \rho_+| \le \begin{cases} (|\rho_- - \rho_+|/2)e^{-\sigma\xi} \le C\delta_M e^{-\sigma\xi} & \text{for} \quad \xi \ge 0, \\ |\rho_- - \rho_+|/2 \le C\delta_M & \text{for} \quad \xi < 0, \end{cases} \tag{2.5b}$$

$$|\tilde{m}(\xi) - m_+| \le \begin{cases} (|m_- - m_+|/2)e^{-\sigma\xi} \le C\delta_M e^{-\sigma\xi} & \text{for} \quad \xi \ge 0, \\ |m_- - m_+|/2 \le C\delta_M & \text{for} \quad \xi < 0. \end{cases} \tag{2.5c}$$

2.2. The equations for the perturbations. The difficulty of the stability problem in the half space comes from the fact that the location of the traveling wave, which should be the asymptotic state for the initial boundary value problem, is shifted owing to the boundary effects. To overcome this difficulty, we employ the idea in [7] and consider the traveling wave in the form of $\tilde{F}(x-st+\alpha-\beta)$, where the parameters, α and β, are to be determined later. Then putting

$$f(x,t) := F(x,t) - \tilde{F}(x - st + \alpha - \beta), \tag{2.6}$$

we have from (1.1) that

$$\partial_t f_1 + v\partial_x f_1 = q(F) - q(\tilde{F}), \tag{2.7a}$$

$$4\partial_t f_2 \qquad\quad = -2(q(F) - q(\tilde{F})), \tag{2.7b}$$

$$\partial_t f_3 - v\partial_x F_3 = q(F) - q(\tilde{F}). \tag{2.7c}$$

The initial and boundary data for (2.7) are given by

$$f_0(x) := f(x,0) = F_0(x) - \tilde{F}(x + \alpha - \beta), \quad f_1(0,t) = B_1 - \tilde{F}_1(-st + \alpha - \beta). \tag{2.8}$$

The local existence theorem for the above system is proved by the standard argument using the characteristic method and the contraction principle.

LEMMA 2.1. *Suppose that $f_0 \in H^1(\mathbb{R}_+)$ and the compatibility condition* (1.5) *holds. Then there exists a positive constant T_0, depending only on $\|f_0\|_1$, such that the initial boundary value problem* (2.7) *and* (2.8) *has a unique solution $f(x,t)$ in the space $C^0([0,T_0];H^1(\mathbb{R}_+)) \cap C^1([0,T_0];L^2(\mathbb{R}_+))$.*

In order to prove the existence of the solution to (2.7) (and thus (1.1)) globally in time, it is convenient to handle the system (1.15) for the fluid dynamical quantities. We regard the solution $(\rho, m, z)(x,t)$ as the perturbation from the traveling wave $(\tilde{\rho}, \tilde{m}, \tilde{z})(x - st + \alpha - \beta)$ and introduce the new unknown functions,

$$\begin{aligned} (\phi,\psi,\omega)(x,t) :&= (\rho, m, z)(x,t) - (\tilde{\rho}, \tilde{m}, \tilde{z})(x - st + \alpha - \beta) \\ &= (f_1 + 4f_2 + f_3, v(f_1 - f_3), v^2(f_1 + f_3))(x,t). \end{aligned} \tag{2.9}$$

From (2.7) (or (1.15)), we have

$$\phi_t + \psi_x = 0, \tag{2.10a}$$

$$\psi_t + \omega_x = 0, \tag{2.10b}$$

$$\omega_t + v^2\psi_x - \tilde{a}\phi - \tilde{m}\psi + \tilde{b}\omega = \Gamma(\phi,\psi,\omega), \tag{2.10c}$$

where the quantities in (2.10c) are given by

$$\tilde{a} := \frac{v^2\tilde{\rho} - \tilde{z}}{4}, \quad \tilde{b} := \frac{v^2\tilde{\rho} + 3\tilde{z}}{4v^2}, \quad \Gamma(\phi,\psi,\omega) := \frac{(v^2\phi - \omega)^2 - 4(\omega^2 - v^2\psi^2)}{8v^2}. \tag{2.11}$$

It holds from (2.5a) with (1.11a) that

$$\tilde{a}' = \frac{1}{4}(v^2 - s^2)\tilde{\rho}' < 0, \quad \tilde{b}' = \frac{1}{4v^2}(v^2 + 3s^2)\tilde{\rho}' < 0. \tag{2.12}$$

Thus, $\tilde{a}$ and $\tilde{b}$ are monotonically decreasing. Using (1.17), (1.11a), and (2.5a), it holds that

$$c \le a^+ \le \tilde{a} \le a^- \le C, \quad c \le b^+ \le \tilde{b} \le b^- \le C \tag{2.13}$$

for certain positive constants c and C, where $a^\pm := \lim_{\eta\to\pm\infty} \tilde{a}(\eta)$, $b^\pm := \lim_{\eta\to\pm\infty} \tilde{b}(\eta)$.

The initial and boundary data for (2.10) are derived from (2.8):

$$(\phi_0, \psi_0, \omega_0)(x) := (\rho_0, m_0, z_0)(x) - (\tilde{\rho}, \tilde{m}, \tilde{z})(x + \alpha - \beta), \tag{2.14}$$

$$\left(\psi + \frac{1}{v}\omega\right)(0,t) = 2v\left(M_1 - \tilde{F}_1(-st + \alpha - \beta)\right). \tag{2.15}$$

We see from (2.10) that we can define the antiderivatives of perturbations ϕ and ψ if $(\phi_0, \psi_0) \in (L^1 \cap H^1)(\mathbb{R}_+)$:

$$\Phi(x,t) := -\int_x^\infty \phi(y,t)\,dy, \quad \Psi(x,t) := -\int_x^\infty \psi(y,t)\,dy. \tag{2.16}$$

Then, integrating (2.10a) and (2.10b) over $[x,\infty)$ for $x > 0$ and multiplying by -1, we have

$$\Phi_t + \Psi_x = 0, \tag{2.17a}$$

$$\Psi_t + \omega = 0, \tag{2.17b}$$

$$\omega_t + v^2\Psi_{xx} - \tilde{a}\Phi_x - \tilde{m}\Psi_x + \tilde{b}\omega = \Gamma(\Phi_x, \Psi_x, \omega). \tag{2.17c}$$

Substituting $\omega = -\Psi_t$ in (2.17c), we obtain the system

$$\Phi_t + \Psi_x = 0, \tag{2.18a}$$

$$\Psi_{tt} - v^2\Psi_{xx} + \tilde{a}\Phi_x + \tilde{m}\Psi_x + \tilde{b}\Psi_t = -\Gamma(\Phi_x, \Psi_x, -\Psi_t). \tag{2.18b}$$

Here, the nonlinear term Γ in (2.18b) is estimated as

$$|\Gamma(\Phi_x, \Psi_x, -\Psi_t)| \le C(|\Phi_x|^2 + |\Psi_x|^2 + |\Psi_t|^2), \tag{2.19a}$$

$$|\partial_t\Gamma(\Phi_x, \Psi_x, -\Psi_t)| \le C(|\Phi_x| + |\Psi_t| + |\Psi_x|)(|\Phi_{xt}| + |\Psi_{tt}| + |\Psi_{xt}|). \tag{2.19b}$$

The initial data for the system (2.18) is derived from (2.14):

$$\Phi_0(x) := -\int_x^\infty \phi_0(y)\,dy, \quad \Psi_0(x) := -\int_x^\infty \psi_0(y)\,dy, \quad \Psi_t(x,0) = -\omega_0(x). \tag{2.20}$$

The boundary condition to be satisfied by (Φ, Ψ) is discussed in the next subsection.

2.3. Boundary estimates. In this subsection, we derive the estimate for the boundary data on $\{x = 0\}$. Divide (2.7b) by 2 and add the resultant equality to (2.7a). This computation yields that

$$(f_1 + 2f_2)_t + v(f_1)_x = 0. \tag{2.21}$$

Integrating (2.21) over $(0,\infty)\times[0,t]$ for $t>0$, we obtain

$$\int_0^\infty (f_1+2f_2)(x,t)\,dx = \int_0^\infty (f_1+2f_2)(x,0)\,dx + v\int_0^t M_1^- - \tilde{F}_1(-s\tau+\alpha-\beta)\,d\tau, \tag{2.22}$$

where we have used (2.8). The left-hand side of (2.22) converges to zero as t tends to infinity, provided that the stability of the traveling wave holds. Thus, it is necessary that the right-hand side of (2.22) converges to zero as t tends to infinity. Therefore, we have

$$\int_0^\infty (f_1+2f_2)(x,0)\,dx + v\int_0^\infty M_1^- - \tilde{F}_1(-s\tau+\alpha-\beta)\,d\tau = 0. \tag{2.23}$$

Thus, it is necessary to chose α so that (2.23) holds. We derive the explicit formula of α following the idea in [7]. We regard the left-hand side of (2.23) as the function of α. Let $I(\alpha)$ denote it. Differentiate $I(\alpha)$ with respect to α using (2.6) and then apply the Rankine–Hugoniot condition in (1.9). The result is $I'(\alpha) = (M_1^- - M_1^+) + 2(M_2^- - M_2^+)$. Integrating this equality in α, we have the identity $I(\alpha) = I(0) + \{(M_1^- - M_1^+) + 2(M_2^- - M_2^+)\}\alpha$. Consequently, we see that (2.23) holds if and only if α is given by

$$\begin{aligned}\alpha &= C_M\left\{\int_0^\infty (F_1+2F_2)(x,0) - (\tilde{F}_1+2\tilde{F}_2)(x-\beta)\,dx \right.\\ &\qquad\left. + v\int_0^\infty M_1^- - \tilde{F}_1(-s\tau-\beta)\,d\tau\right\}\\ &= C_M\left\{-\frac{1}{2}(\hat{\Phi}_0(0)+\frac{1}{v}\hat{\Psi}_0(0)) + v\int_0^\infty M_1^- - \tilde{F}_1(-s\tau-\beta)\,d\tau\right\},\\ C_M &:= \{(M_1^+ - M_1^-) + 2(M_2^+ - M_2^-)\}^{-1}.\end{aligned} \tag{2.24}$$

Notice that the right-hand side of (2.24) is determined by the initial and boundary data. The above observation means that it is necessary to chose α by (2.24) for the stability of the traveling wave. Thus, here and hereafter, we determine α by (2.24) and regard it as the function of the other parameter β. Namely, $\alpha = \alpha(\beta)$. The essential property of $\alpha(\beta)$ is that $\alpha(\beta)$ is bounded even if the parameter β becomes large. Moreover, the following lemma holds.

LEMMA 2.2. *For an arbitrary positive constant ε, there exists a constant δ such that if $\beta>0$ and $\|(\hat{\Phi}_0,\hat{\Psi}_0)\|_1 + 1/\beta \le \delta$, then $|\alpha(\beta)|\le\varepsilon$.*

Proof. Estimating (2.24) by using (2.2), we have

$$|\alpha(\beta)| \le C(|\hat{\Phi}_0|_\infty + |\hat{\Psi}_0|_\infty + e^{-\sigma\beta}). \tag{2.25}$$

The proof immediately follows from (2.25) by applying the Sobolev inequality. ∎

Subtracting (2.23) from (2.22), and using (2.9) and (2.16), we have that

$$\left(\Phi + \frac{1}{v}\Psi\right)(0,t) = B(t), \tag{2.26a}$$

$$B(t) := 2v\int_t^\infty M_1^- - \tilde{F}_1(-s\tau+\alpha(\beta)-\beta)\,d\tau \tag{2.26b}$$

for $t > 0$. Suppose that $\beta > 0$ is sufficiently large. Substitute (2.2) in (2.26b) and apply Lemma 2.2 as well as (2.2). The result is the estimate for boundary data:

$$0 < B(t) < Ce^{-\sigma(st+\beta)} < Ce^{-\sigma\beta}. \tag{2.27}$$

Square (2.27) and integrate in t. The result is

$$0 < \int_0^t B(\tau)^2\,d\tau < C\int_0^t e^{-2\sigma(s\tau+\beta)}\,d\tau < Ce^{-2\sigma\beta}. \tag{2.28}$$

Apply ∂_t^i on (2.26a), for $i = 1, 2$, and use (2.26b). Then we have

$$|\partial_t^i B(t)| < Ce^{-\sigma\beta}, \quad 0 < \int_0^t |\partial_t^i B(\tau)|^2\,d\tau < Ce^{-2\sigma\beta}. \tag{2.29}$$

The existence and the asymptotic state of a time global solution is obtained in the next section by the energy method. In this method, we make the integration by parts, but it gives rise to the integrations in t along the boundary $\{x = 0\}$. These integrations are estimated by utilizing the inequalities (2.27), (2.28), and (2.29).

We conclude this section by showing the next lemma, which we prove using the idea in [7]. Here, we recover a certain gap between the assumptions on initial data in Theorem 1.2 and Proposition 3.1. Precisely, we need to prove that the smallness assumption on the initial data in Theorem 1.2, which is equivalent to the smallness of $(\hat{\Phi}_0, \hat{\Psi}_0, \hat{\omega}_0)$, implies that of $(\Phi_0, \Psi_0, \omega_0)$.

LEMMA 2.3. *For an arbitrary positive constant ε, there exists a positive constant δ such that if $\beta > 0$ and $(\beta+1)(\|(\hat{\Phi}_0, \hat{\Psi}_0)\|_2^2 + \|\hat{\omega}_0\|_1^2) + 1/\beta \le \delta$, then $\|(\Phi_0, \Psi_0)\|_2^2 + \|\omega_0\|_1^2 \le \varepsilon$.*

Proof. Subtracting the first equality in (2.20) from (1.27) yields that

$$\begin{aligned}\hat{\Phi}_0(x) - \Phi_0(x) &= -\int_x^\infty \tilde{\rho}(y-\beta) - \rho_+\,dy + \int_x^\infty \tilde{\rho}(y+\alpha-\beta) - \rho_+\,dy \\ &= \int_\alpha^0 \tilde{\rho}(y+\theta-\beta) - \rho_+\,d\theta =: \chi(x).\end{aligned}$$

Then we compute $\|\chi\|$ with the aid of the estimate (2.5b) assuming $\alpha \ge 0$. The other case, $\alpha < 0$, is computed similarly. Divide the integration region $(0, \infty)$ into three parts and apply (2.5b), respectively. Consequently, we have

$$\begin{aligned}\|\chi\|^2 &= \int_0^{\beta-\alpha} |\chi(x)|^2\,dx + \int_{\beta-\alpha}^{\beta} |\chi(x)|^2\,dx + \int_\beta^\infty |\chi(x)|^2\,dx \\ &\le C(\alpha^2\beta + o(\alpha) + o(1/\beta)),\end{aligned}$$

where $o(\alpha) \to 0$ as $\alpha \to 0$. The first term on the right-hand side of the above inequality is estimated as

$$|\alpha^2\beta| \le C(\|\hat{\Phi}_0\|_1^2 + \|\hat{\Psi}_0\|_1^2 + e^{-2\sigma\beta})|\beta|,$$

where we have used (2.25) and the Sobolev inequality. These computations and Lemma 2.2 yield the estimate

$$\|\Phi_0\|^2 \le \|\hat{\Phi}_0\|^2 + C(\|\hat{\Phi}_0\|_1^2 + \|\hat{\Psi}_0\|_1^2)|\beta| + o(1/\beta).$$

This estimate gives the proof of the assertion concerning $\|\Phi_0\|$. The estimates for the first and the second derivatives of Φ_0 are obtained by direct computation with (2.20) and (1.27). The other terms are handled similarly. □

Note that the assumptions in Lemma 2.3 hold if we choose β sufficiently large and then take the initial data $\|(\hat{\Phi}_0, \hat{\Psi}_0)\|_2 + \|\hat{\omega}_0\|_1$ small enough.

3. Energy estimate. In this section, we prove the a priori estimate in Proposition 3.1 by the energy method following the context of [3]. Theorem 1.2 follows from combining Proposition 3.1 and the local existence of the initial boundary value problem (2.18), (2.20), and (2.26a) with the aid of Lemma 2.3. Precisely, we show the existence of the global solution $(\phi,\psi,\omega)(x,t)$ to (2.10). This immediately means the global existence of the solution $F(x,t)$. The asymptotic convergence (1.28) follows from the standard discussion on the uniform estimate (3.3). (See [7], for example.)

The local existence of the solution $(\phi,\psi,\omega)(x,t)$ to the problem (2.18), (2.20), and (2.26a) follows from Lemma 2.1 and it verifies

$$(\phi,\psi,\omega)\in C^0([0,T_0];H^1(\mathbb{R}_+))\cap C^1([0,T_0];L^2(\mathbb{R}_+)), \tag{3.1}$$

where the existence time T_0 depends only on the initial data $\|(\phi_0,\psi_0,\omega_0)\|_1$. Moreover, if $(\phi_0,\psi_0)\in L^1(\mathbb{R}_+)$, then the antiderivatives $(\Phi,\Psi)(x,t)$ can be defined by (2.16) for $t\in[0,T_0]$ and satisfy (2.18). Furthermore, we see from (2.17) that $(\Phi,\Psi)(\cdot,t)\in L^2(\mathbb{R}_+)$ if $(\Phi_0,\Psi_0)\in L^2(\mathbb{R}_+)$.

We introduce some notations.

$$N(t)=\sup_{0\le\tau\le t}\{\|(\Phi,\Psi)(\tau)\|_2+\|\Psi_t(\tau)\|_1\},$$
$$M(t)^2=\int_0^t\|(\Phi_x,\Phi_{xt},\Psi_x,\Psi_t,\Psi_{xt},\Psi_{tt})(\tau)\|^2\,d\tau,$$
$$|||\phi|||_1:=(\|\phi\|_1^2+\|\phi_t\|^2)^{1/2}.$$

L_1 and L_2 denote the left-hand sides of (2.18), respectively. Namely,

$$L_1(\Phi,\Psi):=\Phi_t+\Psi_x,\quad L_2(\Phi,\Psi):=\Psi_{tt}-v^2\Psi_{xx}+\tilde{a}\Phi_x+\tilde{m}\Psi_x+\tilde{b}\Psi_t. \tag{3.2}$$

Here, we state the a priori estimate. Since the proof is divided in several lemmas, we complete it at the end of this section.

PROPOSITION 3.1. *Let (ϕ,ψ,ω) be a solution to the problem* (2.10), (2.14), *and* (2.15), *which satisfies* (3.1) *with T_0 replaced by T. Then (Φ,Ψ) defined by* (2.16) *is located in $C^0([0,T];H^2(\mathbb{R}_+))$ and satisfies* (2.18), (2.20), *and* (2.26). *Moreover, let δ_M be sufficiently small and then $\beta>0$ be sufficiently large. Then there is a positive constant δ independent of T such that if $N(T)\le\delta$, then we have the uniform estimate*

$$\begin{aligned}&|||(\Phi,\Psi)(t)|||_1^2+\|(\Phi_{xx},\Phi_{xt},\Psi_{xx},\Psi_{xt},\Psi_{tt})(t)\|^2\\&\quad+\int_0^t\|(|\tilde{\rho}'|^{1/2}\Psi)(\tau)\|^2+|||(\Phi_x,\Psi_x,\Psi_t)(\tau)|||_1^2\,d\tau\\&\qquad\le C(\|(\Phi_0,\Psi_0)\|_2+\|\omega_0\|_1^2+e^{-2\sigma\beta})\end{aligned} \tag{3.3}$$

for $0\le t\le T$, where C is a positive constant independent of T.

By the standard argument employing a mollifier with respect to t, we may assume without loss of generality that $(\Phi,\Psi)\in C^\infty([0,T];H^2(\mathbb{R}_+))$.

LEMMA 3.2. *Suppose that the same assumptions as in Proposition* 3.1 *hold. If $N(T)<\delta$, then it holds that*

$$\begin{aligned}&\|\Phi(t)\|^2+|||\Psi(t)|||_1^2\\&\quad+\int_0^t\|(|\tilde{\rho}'|^{1/2}\Psi,\Psi_x,\Psi_t)(\tau)\|^2+\Psi^2(0,\tau)+\Psi_t^2(0,\tau)\,d\tau+\Psi^2(0,t)\\&\qquad\le C(\|\Phi(0)\|^2+|||\Psi(0)|||_1^2+N(t)M(t)^2+e^{-2\sigma\beta})\end{aligned} \tag{3.4}$$

for $0 \le t \le T$, where C is a positive constant independent of T.

Proof. Multiply (2.18a) and (2.18b) by $-\Psi_x$ and $\tilde{a}^{-1}\Psi_t$, respectively. Adding these two resultant equalities yields that

$$-\Psi_x L_1(\Phi,\Psi) + \tilde{a}^{-1}\Psi_t L_2(\Phi,\Psi) = -\tilde{a}^{-1}\Psi_t\Gamma, \tag{3.5}$$

where we have used the notations in (3.2). Precisely, (3.5) gives that

$$\begin{aligned} -\Psi_x^2 + \left\{\tilde{a}^{-1}\left(\frac{1}{2}\Psi_t^2 + \frac{v^2}{2}\Psi_x^2\right) - \Psi_x\Phi\right\}_t + \tilde{a}^{-1}\tilde{b}\Psi_t^2 + \{-v^2\tilde{a}^{-1}\Psi_t\Psi_x + \Psi_t\Phi\}_x \\ - s(\tilde{a}^{-1})'\left(\frac{1}{2}\Psi_t^2 + \frac{v^2}{2}\Psi_x^2\right) + \tilde{a}^{-1}\tilde{m}\Psi_t\Psi_x + (\tilde{a}^{-1})'v^2\Psi_t\Psi_x = -\tilde{a}^{-1}\Psi_t\Gamma. \end{aligned} \tag{3.6}$$

Successively, multiply (2.18a) and (2.18b) by Φ and $\tilde{a}^{-1}\Psi$, respectively, and add the resultant equalities to obtain that

$$\Phi L_1(\Phi,\Psi) + \tilde{a}^{-1}\Psi L_2(\Phi,\Psi) = -\tilde{a}^{-1}\Psi\Gamma. \tag{3.7}$$

Namely,

$$\begin{aligned} \left\{\frac{1}{2}\Phi^2 + \tilde{a}^{-1}\Psi\Psi_t + s(\tilde{a}^{-1})'\frac{1}{2}\Psi^2 + \frac{1}{2}\tilde{a}^{-1}\tilde{b}\Psi^2\right\}_t + \tilde{a}^{-1}(-\Psi_t^2 + v^2\Psi_x^2) \\ + \left\{-\frac{1}{2}\{\tilde{a}^{-1}(\tilde{m} - s\tilde{b})\}' + \frac{1}{2}(s^2 - v^2)(\tilde{a}^{-1})''\right\}\Psi^2 \\ + \left\{\Phi\Psi - v^2\tilde{a}^{-1}\Psi\Psi_x + \frac{1}{2}\tilde{a}^{-1}\tilde{m}\Psi^2 + \frac{v^2}{2}(\tilde{a}^{-1})'\Psi^2\right\}_x = -\tilde{a}^{-1}\Psi\Gamma. \end{aligned} \tag{3.8}$$

Then multiply (3.8) by λ, which is a positive constant to be determined later. Add the resultant equality to (3.6). The result is

$$(E_1 + E_2 + \hat{E}_2)_t + E_3 + E_4 + G + \hat{B}_x = -\tilde{a}^{-1}(\Psi_t + \lambda\Psi)\Gamma, \tag{3.9}$$

where

$$E_1(\Phi,\Psi_x) := \tilde{a}^{-1}\left(\frac{\lambda}{2}\tilde{a}\Phi^2 - \tilde{a}\Phi\Psi_x + \frac{v^2}{2}\Psi_x^2\right), \tag{3.10a}$$

$$E_2(\Psi,\Psi_t) := \tilde{a}^{-1}\left(\frac{\lambda}{2}\tilde{b}\Psi^2 + \lambda\Psi\Psi_t + \frac{1}{2}\Psi_t^2\right), \tag{3.10b}$$

$$\hat{E}_2(\Psi) := \frac{\lambda s}{2}(\tilde{a}^{-1})'\Psi^2, \tag{3.10c}$$

$$E_3(\Psi_t,\Psi_x) := \tilde{a}^{-1}\{(\tilde{b} - \lambda)\Psi_t^2 + \tilde{m}\Psi_t\Psi_x + (\lambda v^2 - \tilde{a})\Psi_x^2\}, \tag{3.10d}$$

$$E_4(\Psi) := \frac{\lambda}{2}\left\{-\{\tilde{a}^{-1}(\tilde{m} - s\tilde{b})\}' + (s^2 - v^2)(\tilde{a}^{-1})''\right\}\Psi^2, \tag{3.10e}$$

$$G(\Psi_t,\Psi_x) := (\tilde{a}^{-1})'\left\{-s\left(\frac{1}{2}\Psi_t^2 + \frac{v^2}{2}\Psi_x^2\right) + v^2\Psi_t\Psi_x\right\}, \tag{3.10f}$$

$$\begin{aligned} \hat{B} := -v^2\tilde{a}^{-1}\Psi_t\Psi_x + \Psi_t\Phi \\ + \lambda\left\{\Phi\Psi - v^2\tilde{a}^{-1}\Psi\Psi_x + \frac{1}{2}\tilde{a}^{-1}\tilde{m}\Psi^2 + \frac{v^2}{2}(\tilde{a}^{-1})'\Psi^2\right\}. \end{aligned} \tag{3.10g}$$

Then, integrating (3.9) over $(0,\infty)\times(0,t)$ we have

$$\begin{aligned}(3.11)\quad &\int_0^\infty (E_1+E_2+\hat{E}_2)(x,t)\,dx\\ &\quad+\int_0^t\int_0^\infty (E_3+E_4+G)(x,\tau)\,dx\,d\tau-\int_0^t \hat{B}(0,\tau)\,d\tau\\ &\quad=\int_0^\infty (E_1+E_2+\hat{E}_2)(x,0)\,dx-\int_0^t\int_0^\infty \tilde{a}^{-1}(\Psi_t+\lambda\Psi)\Gamma\,dx\,d\tau.\end{aligned}$$

Next, we need to estimate the integrands above, respectively. If the shock strength δ_M is sufficiently small, then we have

$$(3.12a)\qquad c(\Phi^2+\Psi_x^2)\le E_1\le C(\Phi^2+\Psi_x^2),$$

$$(3.12b)\qquad c(\Psi^2+\Psi_t^2)\le E_2\le C(\Psi^2+\Psi_t^2),$$

$$(3.12c)\qquad c(\Psi_t^2+\Psi_x^2)\le E_3\le C(\Psi_t^2+\Psi_x^2)$$

by choosing the constant λ suitably, where c and C are positive constants. These equivalences are obtained by computing the determinants of the quadratic forms E_1, E_2, and E_3. For example, if we take λ to be

$$(3.13)\qquad \lambda=\frac{\rho_+-\rho_-}{8}+\frac{z_+-z_-}{8v^2},$$

then (3.12) holds. For details, see [3].

Owing to (2.12) and (2.13), we have

$$(3.14)\qquad 0\le\hat{E}_2(\Psi)\le C|\tilde{\rho}'|\Psi^2,$$

where C is a positive constant. Next, it follows from (2.12) that

$$(3.15)\qquad \{\tilde{a}^{-1}(\tilde{m}-s\tilde{b})\}'=-\frac{v^2-s^2}{4\tilde{a}^2}\rho_\pm(u_\pm-s)\rho'<0,\quad \tilde{a}''=\frac{v^2-s^2}{4}\rho''.$$

The first inequality in (3.15) follows from the inequality

$$(3.16)\qquad -\frac{v^2-s^2}{4\tilde{a}^2}\rho_\pm(u_\pm-s)>c\gneq 0,$$

where c is a positive constant. Here, (3.16) is obtained by using (1.11a). For details, see [3]. Estimating (3.15) using (2.4) and (3.16) and substituting the resultant estimation in (3.10e), we have that

$$(3.17)\qquad E_4\ge c|\tilde{\rho}'|\Psi^2.$$

We also have, from (2.12), that

$$(3.18)\qquad |G|\le C|\tilde{\rho}'|(\Psi_t^2+\Psi_x^2).$$

The last term in (3.11) is estimated by (2.19a) as

$$\begin{aligned}(3.19)\qquad |(\tilde{a}^{-1}(\Psi_t+\lambda\Psi)\Gamma)|&\le C(\|\Psi\|_1+\|\Psi_t\|_1)(\Phi_x^2+\Psi_x^2+\Psi_t^2)\\ &\le C\delta(\Phi_x^2+\Psi_x^2+\Psi_t^2),\end{aligned}$$

where we have used the Sobolev inequality. Substitute the estimates (3.14), (3.17), (3.18), and (3.19) in (3.11). These computations give the desired estimate (3.4) except the boundary integration along $\{x = 0\}$.

Therefore, it remains to estimate the boundary terms. By virtue of (2.26a) and (2.17a), we have

$$\Phi(0,t) = B(t) - \frac{1}{v}\Psi(0,t), \quad \Psi_x(0,t) = -B_t(t) + \frac{1}{v}\Psi_t(0,t). \tag{3.20}$$

Substituting (3.20) in (3.10g) yields that

$$\begin{aligned} -\hat{B}(0,\tau) = \lambda\left\{\frac{1}{v} - \frac{\tilde{m}}{2\tilde{a}} + \frac{v(s-v)}{2}(\tilde{a}^{-1})'\right\}\Psi^2 + \frac{v}{\tilde{a}}\Psi_t^2 \\ + \left\{\left(\frac{1}{2v} + \frac{v\lambda}{2\tilde{a}}\right)\Psi^2\right\}_t - (\lambda\Psi + \Psi_t)(B + v^2\tilde{a}^{-1}B_t), \end{aligned} \tag{3.21}$$

where we have also used (2.13). Then we estimate each term on the right-hand side of (3.21), respectively. The coefficient of Ψ^2 in the first term is strictly positive due to (1.26) with (1.19) if the shock strength δ_M is sufficiently small (see (2.12) and (2.13)). So is the coefficient of Ψ_t^2 in the second term owing to (2.13). We integrate the third term in t to obtain that

$$\left\{\left(\frac{1}{2v} + \frac{v\lambda}{2\tilde{a}}\right)\Psi^2\right\}(0,t) - \left\{\left(\frac{1}{2v} + \frac{v\lambda}{2\tilde{a}}\right)\Psi^2\right\}(0,0). \tag{3.22}$$

Owing to (2.13), the first term in (3.22) appears on the left-hand side of (3.4). The second term is majorized by $\|\Psi_0\|_1$. At last, we estimate the fourth term by using (2.27) and (2.29). The absolute value of the fourth term is less than

$$\varepsilon(|\Psi(0,\tau)|^2 + |\Psi_t(0,\tau)|^2) + C_\varepsilon(|B(\tau)|^2 + |B_t(\tau)|^2), \tag{3.23}$$

where ε is an arbitrary positive number. We take ε so small that the first term in (3.23) is absorbed in the first and the second terms on the right-hand side of (3.21). The second term in (3.23) is estimated by using (2.28) and (2.29). The above computations give the estimates for boundary terms in (2.13) and complete the proof. □

LEMMA 3.3. *Suppose that the same assumptions as in Proposition* 3.1 *hold. If* $N(T) < \delta$*, then it holds that*

$$\begin{aligned} \|\Phi_x(t)\|^2 + \int_0^t \|\Phi_x(\tau)\|^2\,d\tau \\ - c\left\{\|(\Psi_x,\Psi_t)(t)\|^2 + \int_0^t \|(\Psi_x,\Psi_t)(\tau)\|^2 + \Psi_t^2(0,\tau)\,d\tau\right\} \\ \le C(\|(\Phi_x,\Psi_x,\Psi_t)(0)\|^2 + N(t)M(t)^2 + e^{-2\sigma\beta}) \end{aligned} \tag{3.24}$$

for $0 \le t \le T$*, where* c *and* C *are positive constants independent of* T*.*

Proof. Differentiate (2.18a) in x and then multiply by $\Psi_t + v^2\Phi_x$. Then multiply (2.18b) by Φ_x. Adding these two resultant equalities yields that

$$(\Psi_t + v^2\Phi_x)L_1(\Phi_x,\Psi_x) + \Phi_x L_2(\Phi,\Psi) = -\Phi_x\Gamma, \tag{3.25}$$

since $\partial_x L_1(\Phi,\Psi) = L_1(\Phi_x,\Psi_x)$. Precisely, (3.25) gives that

$$\left(\frac{1}{2}v^2\Phi_x^2 + \Psi_t\Phi_x - \frac{1}{2}\Psi_x^2\right)_t + (\Psi_t\Psi_x)_x + \tilde{a}\Phi_x^2 + \Phi_x(\tilde{m}\Psi_x + \tilde{b}\Psi_t) = -\Phi_x\Gamma. \tag{3.26}$$

Then we integrate (3.26) over $(0,\infty)\times(0,t)$ and estimate each term, respectively. The first term on the left-hand side of (3.26) is handled by the Schwarz inequality. The second term gives the boundary integration, which is estimated by the Schwarz inequality as

$$\begin{aligned}\int_0^t |(\Psi_t\Psi_x)(0,\tau)|\,d\tau &\le C\int_0^t B_t^2(\tau) + \Psi_t^2(0,\tau)\,d\tau \\ &\le C(e^{-2\sigma\beta} + \int_0^t \Psi_t^2(0,\tau)\,d\tau),\end{aligned} \tag{3.27}$$

where we have used (3.20) and (2.29). The third term gives the desired term since $\tilde{a} \ge a^+ > 0$ (see (2.13)). The fourth term is handled by the Schwarz inequality again as well as (2.5c) and (2.13):

$$|\Phi_x\Psi_x| \le \varepsilon\Phi_x^2 + C_\varepsilon\Psi_x^2, \quad |\Phi_x\Psi_t| \le \varepsilon\Phi_x^2 + C_\varepsilon\Psi_t^2,$$

where ε is an arbitrary positive constant. We take ε so small that $\varepsilon < a^+$. Finally, the right-hand side of (3.26) is estimated by the argument, using (2.19b), similar to that in deriving (3.19). The above computations yield the desired estimate (3.24). □

Then we need to derive the estimate for the higher derivatives. For this purpose, it is convenient to compute the time derivatives to take advantage of the time decay of the boundary data in (2.27), (2.28), and (2.29). Thus, we differentiate the linear operators $L_1(\Phi,\Psi)$ and $L_2(\Phi,\Psi)$ in t, respectively, to obtain that

$$\partial_t L_1(\Phi,\Psi) = L_1(\Phi_t,\Psi_t), \quad \partial_t L_2(\Phi,\Psi) = L_2(\Phi_t,\Psi_t) - R(\Phi_x,\Psi_x,\Psi_t), \tag{3.28a}$$

$$R(\Phi_x,\Psi_x,\Psi_t) := s(\tilde{a}'\Phi_x + \tilde{m}'\Psi_x + \tilde{b}'\Psi_t). \tag{3.28b}$$

LEMMA 3.4. *Suppose that the same assumptions as in Proposition* 3.1 *hold. If* $N(T) < \delta$, *then it holds that*

$$\begin{aligned}&\|\Phi_t(t)\|^2 + |||\Psi_t(t)|||_1^2 + \int_0^t \|(\Psi_{xt},\Psi_{tt})(\tau)\|^2 + \Psi_t^2(0,\tau) + \Psi_{tt}^2(0,\tau)\,d\tau \\ &\qquad + \Psi_t^2(0,t) - c\int_0^t \|(\Phi_x,\Psi_x,\Psi_t)(\tau)\|^2\,d\tau \\ &\qquad\qquad \le C(\|\Phi_t(0)\|^2 + |||\Psi_t(0)|||_1^2 + N(t)M(t)^2 + e^{-2\sigma\beta})\end{aligned} \tag{3.29}$$

for $0 \le t \le T$, *where* c *and* C *are positive constants independent of* T.

Proof. We first differentiate (2.18a) in t and multiply by $-\Psi_{xt}$. Successively, differentiate (2.18b) in t and multiply by $\tilde{a}^{-1}\Psi_{tt}$. Adding these two resultant equalities using (3.28), we have

$$-\Psi_{xt}L_1(\Phi_t,\Psi_t) + \tilde{a}^{-1}\Psi_{tt}L_2(\Phi_t,\Psi_t) = -\tilde{a}\Psi_{tt}\Gamma_t + \tilde{a}^{-1}\Psi_{tt}R. \tag{3.30}$$

Next, differentiate (2.18a) in t and multiply by Φ_t. Then, differentiate (2.18b) in t and multiply $\tilde{a}^{-1}\Psi_t$. Adding these two equalities yields that

$$\Phi_t L_1(\Phi_t,\Psi_t) + \tilde{a}^{-1}\Psi_t L_2(\Phi_t,\Psi_t) = -\tilde{a}^{-1}\Psi_t\Gamma_t + \tilde{a}^{-1}\Psi_t R. \tag{3.31}$$

It is easy to see that the left-hand sides of (3.30) and (3.31) take the forms of the left-hand sides of (3.5) and (3.7) with (Φ_t, Ψ_t) in place of (Φ, Ψ), respectively. Thus, we apply the computation in deriving (3.4) to the left-hand sides of (3.30) and (3.31). Since the nonlinear terms including Γ_t in (3.30) and (3.31) are estimated, using (2.19b), by the same method as (3.19), it suffices to estimate the last terms on the right-hand sides of (3.30) and (3.31). In fact, they are handled by using the inequalities

$$|\tilde{a}^{-1}\Psi_{tt}R| \le \varepsilon\Psi_{tt}^2 + C_\varepsilon(\Phi_x^2 + \Psi_x^2 + \Psi_t^2), \quad |\tilde{a}^{-1}\Psi_t R| \le C(\Phi_x^2 + \Psi_x^2 + \Psi_t^2), \tag{3.32}$$

where ε is an arbitrarily small positive constant. In deriving (3.32), we have used (2.5a) and (2.12). Consequently, we have the desired estimate (3.29). □

LEMMA 3.5. *Suppose that the same assumptions as in Proposition* 3.1 *hold. If* $N(T) < \delta$, *then it holds that*

$$\begin{aligned} &\|(\Phi_{xt}, \Psi_{xx})(t)\|^2 + \int_0^t \|(\Phi_{xt}, \Psi_{xx})(\tau)\|^2\, d\tau \\ &\quad - c\left\{\|(\Psi_{xt}, \Psi_{tt})(t)\|^2 + \int_0^t \|(\Phi_x, \Psi_x, \Psi_t, \Psi_{xt}, \Psi_{tt})(\tau)\|^2 + \Psi_{tt}^2(0,\tau)\, d\tau\right\} \\ &\qquad\qquad \le C(\|\Phi_{xt}, \Psi_{xt}, \Psi_{tt})(0)\|^2 + N(t)M(t)^2) \end{aligned} \tag{3.33}$$

for $0 \le t \le T$, *where* c *and* C *are positive constants independent of* T.

Proof. Differentiate (2.18) with respect to t, respectively. Then, apply the procedure in deriving (3.25). The result is

$$(\Psi_{tt} + v^2\Phi_{xt})L_1(\Phi_{xt}, \Psi_{xt}) + \Phi_{xt}L_2(\Phi_t, \Psi_t) = -\Phi_{xt}\Gamma_t + \Phi_{xt}R. \tag{3.34}$$

Here, notice that the left-hand side of (3.34) takes the form of the left-hand side of (3.25) with (Φ_t, Ψ_t) in place of (Φ, Ψ). Thus, the left-hand side of (3.34) is handled by the same procedure as that in the proof of Lemma 3.3. Since the first term on the right-hand side of (3.34) is estimated by using (2.19b) with the Sobolev and the Schwarz inequalities, it suffices to drive the estimation on the last term in (3.34). It is handled by applying the Schwarz inequality on (3.28b) as

$$|\Phi_{xt}R| \le \varepsilon\Phi_{xt}^2 + C_\varepsilon(\Phi_x^2 + \Psi_x^2 + \Phi_t^2),$$

where ε is an arbitrarily small positive constant. Finally, use the equality $|\Phi_{tx}| = |\Psi_{xx}|$, which follows from (2.18a). Consequently, we have the desired estimate (3.33). □

In order to complete the proof of the a priori estimate (3.3), we derive the estimates for the other remaining terms.

LEMMA 3.6. *Suppose that the same assumptions as in Proposition* 3.1 *hold. If* $N(T) < \delta$, *then it holds that*

$$\begin{aligned} &\|\Phi_{xx}(t)\|^2 + \int_0^t \|\Phi_{xx}(\tau)\|^2\, d\tau \\ &\quad - c\left\{\|\Psi_{xt}(t)\|^2 + \int_0^t \|(\Phi_x, \Psi_x, \Psi_t, \Phi_{xt}, \Psi_{xt})(\tau)\|^2\, d\tau\right\} \le C\|(\Phi_{xx}, \Psi_{xt})(0)\|^2 \end{aligned} \tag{3.35}$$

for $0 \le t \le T$, *where* c *and* C *are positive constants independent of* T.

Proof. Differentiating (2.7b) in x, we have

$$f_{2\,tx} + \tilde{F}_2 f_{2\,x} = -(f_{2\,x} + \tilde{F}_{2\,x}) f_2 + H, \tag{3.36a}$$

$$H := \frac{1}{2}(f_1 f_3 + \tilde{F}_3 f_1 + \tilde{F}_1 f_3 + \tilde{F}_1 f_{3\,x})_x. \tag{3.36b}$$

Multiplying (3.36a) by $f_{2\,x}$ yields that

$$\frac{1}{2}(f_{2\,x}^2)_t + \tilde{F}_2 f_{2\,x}^2 = -(f_{2\,x} + \tilde{F}_{2\,x}) f_2 f_{2\,x} + H f_{2\,x} \tag{3.37a}$$

$$\leq (c_1\delta + \varepsilon) f_{2\,x}^2 + C_\varepsilon (f_2^2 + H^2), \tag{3.37b}$$

where ε is an arbitrary positive constant and c_1 is a certain positive constant. In deriving the inequality (3.37b), we have used the Schwarz inequality and the fact that

$$|f_2|_\infty \leq c_1(\|\Phi_x\|_1 + \|\Psi_t\|_1) \leq c_1\delta, \tag{3.38}$$

which follows from (2.9). In addition, we have from (2.9) that

$$f_2^2 \leq C(\Phi_x^2 + \Psi_t^2), \quad H^2 \leq C(\Psi_x^2 + \Psi_t^2 + \Psi_{xx}^2 + \Psi_{xt}^2). \tag{3.39}$$

Substitute the estimates (3.39) in (3.37b), take ε and δ so small that $c_1\delta + \varepsilon < M_2^+ = \min_{\xi\in\mathbb{R}} \tilde{F}_2(\xi)$, and successively integrate the resultant inequality to obtain that

$$\|f_{2\,x}(t)\|^2 + \int_0^t \|f_{2\,x}(\tau)\|^2\, d\tau \leq C\left\{\|f_{2\,x}(0)\|^2 + \int_0^t \|(\Phi_x, \Psi_x, \Psi_t, \Psi_{xx}, \Psi_{tx})(\tau)\|^2\, dt\right\}. \tag{3.40}$$

From (2.9), we have

$$c(|\Phi_{xx}|^2 - |\Psi_{xt}|^2) \leq |f_{2\,x}|^2 \leq C(|\Phi_{xx}|^2 + |\Psi_{xt}|^2), \tag{3.41}$$

where c and C are positive constants. Substituting (3.41) in (3.40) yields the desired estimate (3.35). ∎

Proof of Proposition 3.1. Multiply the estimates (3.4), (3.24), (3.29), (3.33), and (3.35) by suitably chosen positive constants, respectively, sum up the resultant inequalities, and then take δ sufficiently small. Consequently, it holds that

$$\begin{aligned} &|||(\Phi,\Psi)(t)|||_1^2 + \|(\Phi_{xx}, \Phi_{xt}, \Psi_{xx}, \Psi_{xt}, \Psi_{tt})(t)\|^2 \\ &\quad + \int_0^t \|(|\tilde{\rho}'|^{1/2}\Psi)(\tau)\|^2 + |||(\Phi_x, \Psi_x, \Psi_t)(\tau)|||_1^2\, d\tau \\ &\quad \leq C(|||(\Phi,\Psi)(0)|||_1^2 + \|(\Phi_{xt}, \Phi_{xx}, \Psi_{xx}, \Psi_{xt}, \Psi_{tt})(0)\|^2 + e^{-2\sigma\beta}). \end{aligned} \tag{3.42}$$

From (2.17) and (2.18), we see that the right-hand side of (3.42) is equivalent to $(\|\Phi_0, \Psi_0)\|_2^2 + \|\omega_0\|_1^2 + e^{-2\sigma\beta})$. These procedures yield the desired estimate (3.3). ∎

Acknowledgments. The author would like to express his deepest gratitude to Professor Shuichi Kawashima and Professor Akitaka Matsumura for stimulating discussions and helpful comments.

REFERENCES

[1] C. Bose, R. Illner, and S. Ukai, *On shock wave solutions for discrete velocity models of the Boltzmann equation*, Transport Theory Statist. Phys., 27 (1998), pp. 35–66.

[2] R. E. Caflish, *Navier-Stokes and Boltzmann shock profiles for a model of gas dynamics*, Comm. Pure Appl. Math., 32 (1979), pp. 521–554.

[3] S. Kawashima and A. Matsumura, *Asymptotic stability of traveling wave solution of systems for one-dimensional gas motion*, Comm. Math. Phys., 101 (1985), pp. 97–127.

[4] S. Kawashima and S. Nishibata, *Existence of a stationary wave for the discrete Boltzmann equation in the half space*, Comm. Math. Phys., 207 (1999), pp. 385–409.

[5] S. Kawashima and S. Nishibata, *A stationary wave for the discrete Boltzmann equation with the reflective boundary*, Comm. Math. Phys., 211 (2000), pp. 167–182.

[6] J.-G. Liu and Z. Xin, *Boundary-layer behavior in the fluid-dynamic limit for a nonlinear model Boltzmann equation*, Arch. Ration. Mech. Anal., 135 (1996), pp. 61–105.

[7] A. Matsumura and M. Mei, *Asymptotics toward viscous shock profile for solution of the viscous p-system with boundary effect*, Arch. Ration. Mech. Anal., 146 (1999), pp. 1–22.

[8] Y. Nikkuni and S. Kawashima, *Stability of stationary solutions to the half-space problem for the discrete Boltzmann equation with multiple collisions*, Kyushu J. Math., 54 (2000), pp. 233–255.

[9] Y. Nikkuni and S. Kawashima, *Asymptotic stability of rarefaction waves for some discrete velocity model of the Boltzmann equation in the half-space*, Adv. Math. Sci. Appl., 12 (2002), pp. 327–353.

[10] S. Ukai, *On the half-space problem for discrete velocity model of the Boltzmann equation*, in Advances in Nonlinear Partial Differential Equations and Stochastics, Ser. Adv. Math. Appl. Sci. 48, S. Kawashima and T. Yanagisawa, eds., World Scientific, Singapore, 1998, pp. 160–174.

SIAM J. MATH. ANAL.
Vol. 34, No. 3, pp. 573–594

TIME DISCRETIZATION OF TRANSITION LAYER DYNAMICS IN ONE-DIMENSIONAL VISCOELASTIC SYSTEMS*

H. LIM†

Abstract. We investigate how evolution occurs as the strain u_x of a viscoelastic system $u_{tt} - (\sigma(u_x) + u_{xt})_x + u = 0$ goes towards a state of equilibrium. The time limit of u_x eventually shows a finite number of discontinuous interfaces if the strain starts from the continuous initial data whose transition layers are steep enough and the initial energy is sufficiently small. The number of phases is conserved and the transition layers stay in the initial position of interfaces. The results are obtained by using the implicit time discretization method and the Andrews–Pego transformed equations.

Key words. implicit time discretization, viscous dissipation, transition layers, Andrews–Pego transformed equations, nonconvex energy

AMS subject classifications. 35G25, 74N25, 74D10

PII. S0036141001398770

Introduction. There are various results on the phase transitions of microstructured elastic crystals [2, 5, 12, 16, 25, 27, 30, 32, 33]. Nonconvex double-well free energy induces hysteretic behavior of the fine microstructures of the material. The usual approach involves the minimization of the elastic energy. Due to the nonconvexity of the free energy, every minimizing sequence fails to attain a minimizer and induces the formation of finer and finer oscillations of the sequence [5, 6, 30]. However, under the presence of the energy dissipation, such a behavior is prevented and the solution converges to the minimizer of the energy [3, 15, 27].

This article focuses on the one-dimensional viscoelastic system

$$u_{tt} - (\sigma(u_x) + u_{xt})_x + u = 0, \tag{0.1}$$

where u is a mapping from $(0,1) \times (0,\infty) \subset \mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$ under appropriate initial and boundary conditions and $\sigma(x) = W'(x)$ for some stored energy function $W : \mathbb{R} \to \mathbb{R}$.

The system describes a time-dependent elastic bar with a nonconvex energy W and a viscous stress u_{xxt} with the zero displacement boundary conditions. The bar interacts with an elastic foundation u. In other words, the bar is placed on a system of linearly elastic springs [32].

Many global existence results for the solutions of similar systems are available [1, 3, 4, 7, 9, 11, 13, 14, 15, 16, 17, 18, 21, 24, 26, 27, 28]. The existence of a weak solution for the viscoelastic-type materials was developed for the cases without assuming the ellipticity of the free energy W [27], the convexity of W, or the Lipschitz continuity of σ [15]. In either case, the viscous dissipation plays a significant role in the strong convergence of the minimizing sequences. In the higher-dimensional case, Friesecke and Dolzmann [15] approached the results by an approximation, called the time discretized solution, on each sufficiently small time interval and using the Andrews–Pego transformed equations which were introduced in [1, 25].

The dynamics of the transition layers on the viscoelastic system (0.1) is the main topic in this paper. The transition layers are defined by the part of the strain u_x

*Received by the editors November 27, 2001; accepted for publication (in revised form) April 29, 2002; published electronically December 13, 2002.
http://www.siam.org/journals/sima/34-3/39877.html
†Department of Mathematics, Michigan State University, East Lansing, MI 48824 (lim@math.msu.edu).

where the norm of u_x is less than a sufficiently small number. For the dynamics, we need the following two assumptions:

(I) The continuous initial data u_0 must have steep enough transition layers; that is, the norm of $(u_0)_{xx}$ should be sufficiently large at the position of transition layers of the initial data.

(II) The initial energy where the energy functional is defined by

$$E(u,v) = \int_0^1 \left[\frac{1}{2}(u(x))^2 + W((u_x(x)) + \frac{1}{2}(v(x))^2\right] dx$$

must be sufficiently small.

Under these assumptions, the time limit of u_x experiences a discontinuity at a finite number of points. More precisely, the transition layers get steeper and eventually become discontinuous at the time limit. Away from these finitely many points, the solution remains continuous and converges to a steady state. The number of transition layers remains the same and the transition layers of the solution are always within the intervals of initial layers, which is comparable to the results of stick-slip motion of layers in a system with the time-dependent displacement boundary conditions [33]. It was proven in [33] that the dynamics exhibits a different behavior than our main results. The layers do not stay in the initial intervals and will move both forward and backward.

Some important physical applications include a phase transformation in microstructured elastic crystals caused by changes in temperature, stresses, or incident electromagnetic waves. The nonconvex elastic energy functional induces finer and finer oscillations, but under the given initial state, the dynamics prevents the nucleation of more phases.

Friesecke and McLeod [16] proved the dynamics of transition layers employing the semigroup approach. In this paper, we use the method of time discretization to show the results. The time discretization theory has a long history [8, 10, 15, 19]. The scheme has been used for the nonlinear diffusion equations [8] and for the parabolic equations [19]. The second order time discretization on a related problem was first introduced in [10].

The existence of the solution was achieved in [16] by proving the existence of the Andrews–Pego transformations under an appropriate fractional power space. However, in [15], the variational approach was utilized for the proof of existence. It was proven that the time discretized functional

$$J^{m,j}[u] := \int_0^1 \left[\frac{1}{2m^2}|u - 2u^{m,j-1} + u^{m,j-2}|^2 + W(u_x) + \frac{1}{2m}|u_x - u_x^{m,j-1}|^2 + \frac{1}{2}|u|^2\right] dx$$

for each time interval $((j-1)m,\ jm],\ \ j \in \mathbb{N}$, where $m > 0$ is a fixed and sufficiently small time stepsize, has a minimizer $u^{m,j}$ which is a weak solution to the following discretized version of (0.1):

$$\frac{1}{m^2}(u - 2u^{m,j-1} + u^{m,j-2}) - (\sigma(u_x))_x - \frac{1}{m}(u_x - u_x^{m,j-1})_x + u = 0 \tag{0.2}$$

on $((j-1)m,\ jm],\ \ j \in \mathbb{N}$. The key idea of the approach is the following: The nonconvexity of the stored energy function W is compensated by the discretized viscous damping term $\frac{1}{2m}|u_x - u_x^{m,j-1}|^2$ to provide the convexity of the functional. It was also proven in [15] that the $W_0^{1,p}$ limit of the time discretized solutions is the weak

solution of (0.1) as the time stepsize m approaches zero. The time discretization scheme naturally applies to the proof of the dynamics of the phase transition in this paper. The method is rather straightforward since it does not introduce any new space. There is nevertheless still a question whether the asymptotic behaviors of (0.1) and (0.2) commute or not. Is there any equivalence between $jm \to \infty$ for fixed m and then $m \to 0$ for the discretized problem, and $t \to \infty$ for the original problem? The question is discussed in the last section of the paper.

As in previous works [16, 27], the decay of the energy functional $E(u, v)$ is the crucial point of the proof. We prove in section 4 that the discretized energy functional is nonincreasing and bounded by the initial energy. A priori estimates are also proved in this section. We show next the existence and the equilibrium state as the limit $u_\star^m$ (as $j \to \infty$) of the discretized solution for fixed m in section 5. The main proof of the dynamics is conducted in section 6. We prove in this section that the transition layers approach a jump discontinuity as $j \to \infty$ by showing that a finite number of intervals where the strain is steep enough are decreasing to a finite number of isolated points. Unfortunately, the intervals in $(0, 1)$ where the norm of the strain $u_x^{m,j}$ is sufficiently small (denote the intervals as $I(u_x^{m,j})$) do not decrease monotonically as $j \to \infty$ in general. Instead, we introduce the time discretized version of Andrews–Pego transformed equations,

$$p^{m,j}(x) := \frac{1}{m}\int_0^x [u^{m,j}(y) - u^{m,j-1}(y)]dy - \frac{1}{m}\int_0^1 \int_0^z [u^{m,j}(y) - u^{m,j-1}(y)]dydz,$$
$$q^{m,j}(x) := u_x^{m,j}(x) - p^{m,j}(x),$$

and consider the finite number of intervals in $(0, 1)$ where the norm of $q^{m,j}$ is sufficiently small (denote them as $I(q^{m,j})$). We show that the $I(q^{m,j})$ decrease monotonically and exponentially to the isolated points as $j \to \infty$, and the intervals $I(u_x^{m,j})$ are contained in the $I(q^{m,j})$. The solution at the limit exhibits a jump discontinuity because of the decrease of the $I(q^{m,j})$ and the fact that the $I(u_x^{m,j})$ are contained in the $I(q^{m,j})$. In the last section, we summarize the relationship between the asymptotic behaviors of (0.1) and (0.2).

The interaction of the bar with an elastic foundation u induces a finely layered microstructure [12]. It has also been shown using the bifurcation analysis that the elastic foundation induces oscillations in the one-dimensional case of the static problem [32]. Nevertheless, under the assumption of low initial energy, the results still hold without the elastic foundation, and only minor change of the proof is required.

Another advantage is that the method is also useful in the practical implementation of the numerical solution of the system. The numerical results for the related problems have been discussed [4, 20, 29, 30, 31]. An implicit finite difference scheme for the homogeneous boundary conditions was achieved in [29], and the numerical methods for the time-dependent boundary conditions were investigated in [31]. In [20], the efficient numerical algorithms for the system in both one- and two-dimensional cases were developed. Applications to the microscale heat transfer equations will also appear in the near future.

The dynamics on similar problems was investigated in [22, 23]. The stability of the incompressible viscoelastic non-Newtonian fluid flows was observed in these papers. Investigating this type of spurt phenomena using the method of time discretization would be very interesting for future work.

1. The initial-boundary value problem and hypotheses. Consider the initial-boundary value problem

$$
\begin{aligned}
&u_{tt} - (\sigma(u_x) + u_{xt})_x + u = 0 \qquad (x \in (0,1), \quad t \in (0,\infty)),\\
&u|_{x=0} = u|_{x=1} = 0 \qquad (t \in [0,\infty)),\\
&u|_{t=0} = u_0, \quad u_t|_{t=0} = v_0 \qquad (x \in [0,1]),
\end{aligned} \tag{1.1}
$$

where u is a function from $(0,1)\times(0,\infty) \subset \mathbb{R}\times\mathbb{R}$ to $\mathbb{R}$, $\sigma = W^{'}$, and W is a stored energy function satisfying the following conditions:

(H1) $W \in C^2(\mathbb{R}), W^{'} = \sigma$.

(H2) There exist $c > 0,\ C > 0$, and $p \geq 2$ such that $c|z|^p - C \leq W(z) \leq C(|z|^p + 1),\ \ |\sigma(z)| \leq C(|z|^{p-1} + 1)$ (coercivity).

(H3) W: double-well potential, that is, there exist $z_- < z_1 < 0 < z_2 < z_+$ such that $W(z_\pm) = 0, W > 0$ elsewhere, $W^{'}(0) = 0, W^{''}|_{(z_1,z_2)} < 0, W^{''}|_{\mathbb{R}\setminus[z_1,z_2]} > 0$.

The stored energy function W is usually a fourth order nonconvex polynomial, and the most common example is $W(z) = \frac{1}{4}(z^2-1)^2$. Moreover, assume

(A1) (smoothness and a priori bounds) $u_0 \in C^2, v_0 \in W_0^{1,2}, ||(u_0)_x||_{L^\infty} + ||v_0||_{W^{1,2}} \leq M$;

(A2) (low initial energy) $E(u_0, v_0) < \epsilon$, where

$$E(u,v) := \int_0^1 \left(\frac{1}{2}u^2 + W(u_x) + \frac{1}{2}v^2\right) dx;$$

(A3) (no transition layers at $x = 0, 1$) $\mathcal{L}_\rho(0) := \{x \in [0,1] : |(u_0)_x(x)| \leq \rho\} \subset (0,1)$;

(A4) (steepness of transition layers) $|(u_0)_{xx}(x)| \geq K$ in $\mathcal{L}_\rho(0)$

for some $M, \epsilon, \rho, K > 0$. Here, $\epsilon,\ \rho$ are sufficiently small numbers and K is a sufficiently large number.

Let the connected components of $\mathcal{L}_\rho(0)$ be denoted by $[(a_0)_i, (b_0)_i]$, $i = 1, \ldots, N$ $(0 < (a_0)_1 < (b_0)_1 < \cdots < (a_0)_N < (b_0)_N < 1)$. Note that by assumption (A4), there exists only one zero of $(u_0)_x$ in each interval $[(a_0)_i, (b_0)_i]$. Let the zeros of $(u_0)_x$ be $(x_0)_i, (x_0)_i \in [(a_0)_i, (b_0)_i], i = 1, \ldots, N$.

2. The time discrete scheme for the solution. Let $m > 0,\ m \ll 1$ be the time stepsize of our problem. The m will be fixed throughout the paper except for the last section. Let $u^{m,0} := u_0, v^{m,0} := v_0, u^{m,-1} := u_0 - mv_0$. For each time interval $((j-1)m,\ jm], j \in \mathbb{N}$, define the following functional inductively:

$$J^{m,j}[u] := \int_0^1 \left[\frac{1}{2m^2}|u - 2u^{m,j-1} + u^{m,j-2}|^2 + W(u_x) + \frac{1}{2m}|u_x - u_x^{m,j-1}|^2 + \frac{1}{2}|u|^2\right] dx$$

on the Sobolev space $W_0^{1,p}((0,1),\mathbb{R})$, where p is the coercivity exponent of W in (H2). It was shown that $J^{m,j}$ attains a minimum $u^{m,j}$ if W satisfies the hypotheses (H1), (H2), and (H3) since the first and the fourth integrands are convex and the nonconvex term $W(u_x)$ combined with the viscous dissipation term $\frac{1}{2m}|u_x - u_x^{m,j-1}|^2$ provides the weakly lower semicontinuity [15]. It can be easily verified that for each time interval $((j-1)m,\ jm],\ j \in \mathbb{N}$, $u^{m,j}(x)$ is the weak solution of

$$\frac{1}{m^2}(u - 2u^{m,j-1} + u^{m,j-2}) - (\sigma(u_x))_x - \frac{1}{m}(u_x - u_x^{m,j-1})_x + u = 0, \tag{2.1}$$

which is the time approximated equation of (1.1). The $u^{m,j}$ is thus called the time discretized solution of the problem (1.1). Assume that $u^{m,j}$ satisfies the boundary

conditions of (1.1) for each $j \in \mathbb{N}$. On the time interval $((j-1)m,\ jm],\ j \in \mathbb{N}$, we define the linear interpolation function $u^j(x,t)$ of $u^{m,j}(x)$ as follows:

$$u^j(x,t) := \left(\frac{mj-t}{m}\right) u^{m,j-1}(x) + \left(\frac{t-m(j-1)}{m}\right) u^{m,j}(x). \tag{2.2}$$

It is now important to define the functions which are called the Andrews–Pego transformed equations. The equations will play a crucial role for the proof of the main results. Define

$$\begin{aligned}
p_0(x) &:= \int_0^x v_0(y)dy - \int_0^1 \int_0^z v_0(y)dydz,\\
q_0(x) &:= (u_0)_x(x) - p_0(x),\\
p^{m,j}(x) &:= \frac{1}{m}\int_0^x [u^{m,j}(y) - u^{m,j-1}(y)]dy - \frac{1}{m}\int_0^1 \int_0^z [u^{m,j}(y) - u^{m,j-1}(y)]dydz,\\
q^{m,j}(x) &:= u_x^{m,j}(x) - p^{m,j}(x)
\end{aligned}$$

for all $j \in \mathbb{N}$. Note that $p_x^{m,j}(x) = \frac{u^{m,j}(x)-u^{m,j-1}(x)}{m} =: v^{m,j}(x)$. For all $j \in \mathbb{N}$ and $(j-1)m < t \le jm$, define the interpolation functions of $p^j(x,t),\ q^j(x,t)$, and $v^j(x,t)$ of $p^{m,j}(x),\ q^{m,j}(x)$, and $v^{m,j}(x)$, respectively, in the same way as (2.2).

3. Main results. The following theorem describes the dynamical behavior of the transition layers.

THEOREM 3.1. *Suppose the stored energy function W and the initial data $(u_0,\ v_0) \in W_0^{1,\infty} \times L^2$ are assumed to satisfy* (H1)–(H3), (A1)–(A4). *Then the following hold:*

(P1) (Preservation of number of zeros.) *The number of zeros of $u_x^{m,j}$, denoted by $N(j)$, is finite, is positive for all $j \in \{0\} \cup \mathbb{N}$, and is independent of j. Let the zeros be denoted by $0 < x_1^m(j) < \cdots < x_N^m(j) < 1$.*

(P2) (Preservation of intervals of transition layers.) *The number of connected components of $\mathcal{L}_{\frac{\rho}{2}}(j) := \{x \in (0,1) : |u_x^j(x,t)| \le \frac{\rho}{2}\}$ is finite, is positive for all $j \in \mathbb{N}$, and is independent of j, and in each connected component, $u_x^j(x,t)$ is strictly monotone and has exactly one zero. Let the connected components of $\mathcal{L}_{\frac{\rho}{2}}(j)$ be denoted by $[a_i^m(j), b_i^m(j)],\ i = 1, \dots, N$ $(0 < a_1^m(j) < b_1^m(j) < \cdots < a_N^m(j) < b_N^m(j) < 1)$.*

(P3) (Lock-in and exponential steepening of transition layers.) *For all $j \in \mathbb{N}$, $i = 1, \dots, N$, and for some $K_0 > 0$,*

$$x_i^m(j) \in [a_i^m(j), b_i^m(j)] \subset [(a_0)_i, (b_0)_i],$$

$$|u_{xx}^j(x,t)| \ge \frac{K_0}{2} e^{\sigma_0 jm} \quad \forall x \in \mathcal{L}_{\frac{\rho}{2}}(j) = \bigcup_{i=1}^N [a_i^m(j), b_i^m(j)],$$

$$|b_i^m(j) - a_i^m(j)| \le \frac{2\rho}{K_0} e^{-\sigma_0 jm},$$

where $\sigma_0 := \min_{[-\rho,\rho]} |\sigma'| > 0$.

(P4) (Convergence of phases.) $\lim_{j\to\infty} x_i^m(j) =: (x_\star)_i^m$ *exists for all $i = 1, \dots, N$ and $(x_\star)_i^m \in [(a_0)_i, (b_0)_i]$ (in particular, $0 < (x_\star)_1^m < \cdots < (x_\star)_N^m < 1$).*

(P5) (Jump discontinuity of the limit state.) $\lim_{j\to\infty} u_x^{m,j} =: (u_\star^m)_x$ *(which exists as an L^p limit) is continuous on $(0,1)\backslash\{(x_\star)_1^m, \dots, (x_\star)_N^m\}$ but discontinuous at every $(x_\star)_i^m$ for all $i = 1, \dots, N$.*

4. Energy decay and a priori estimates. Let $t \in ((j-1)m,\ jm],\ j \in \mathbb{N}$. We first prove the decay of the energy functional:

$$E(t) = E(u^j, v^j) = \int_0^1 \left[\frac{1}{2}(u^j(x,t))^2 + W(u_x^j(x,t)) + \frac{1}{2}(v^j(x,t))^2\right] dx.$$

LEMMA 4.1. *$E(t)$ is nonincreasing, bounded by the initial data on $((j-1)m,\ jm]$ for all $j \in \mathbb{N}$.*

Proof. Recall that $u^{m,j},\ j \in \mathbb{N}$, satisfies (2.1). That is, the equation

$$v_t^j - \sigma(u_x^{m,j})_x - v_{xx}^{m,j} + u^{m,j} = 0 \tag{4.1}$$

is satisfied for all $j \in \mathbb{N}$. Then the following holds:

$$\begin{aligned}\frac{d}{dt}E(t) = &\int_0^1 [v^{m,j}\cdot v_t^j + \sigma(u_x^{m,j})\cdot u_{xt}^j + u^{m,j}\cdot u_t^j + (v^j - v^{m,j})v_t^j \\ &+ (\sigma(u_x^j) - \sigma(u_x^{m,j}))\cdot u_{xt}^j + (u^j - u^{m,j})\cdot u_t^j]dx \qquad (4.2)\\ = &\int_0^1 [v^{m,j}\{v_t^j - \sigma(u_x^{m,j})_x + u^{m,j}\} + (v^j - v^{m,j})v_t^j \\ &+ (\sigma(u_x^j) - \sigma(u_x^{m,j}))\cdot u_{xt}^j + (u^j - u^{m,j})\cdot u_t^j]dx \qquad (4.3)\\ = &\int_0^1 \left[v^{m,j}\cdot v_{xx}^{m,j} + \frac{(t-jm)}{m^2}\cdot|v^{m,j} - v^{m,j-1}|^2\right. \\ &\left. + (\sigma(u_x^j) - \sigma(u_x^{m,j}))\cdot u_{xt}^j + \frac{(t-jm)}{m^2}\cdot|u^{m,j} - u^{m,j-1}|^2\right] dx \\ = &-\int_0^1 |v_x^{m,j}|^2 dx + \frac{(t-jm)}{m^2}\int_0^1 |v^{m,j} - v^{m,j-1}|^2 dx \\ &+ \int_0^1 (\sigma(u_x^j) - \sigma(u_x^{m,j}))\cdot u_{xt}^j dx + (t-jm)\int_0^1 |v^{m,j}|^2 dx \qquad (4.4)\end{aligned}$$

for $(j-1)m < t \le jm$. The first three terms of (4.2) are the same as the first term of (4.3) by the integration by parts and the boundary conditions of (1.1). By using the mean value theorem, and by the fact that the function σ' is bounded below by a negative number, that is, for all $y \in \mathbb{R}$, $\sigma'(y) \ge -L$ for some $L > 0$, the integrand of the third term of (4.4) is estimated in the following way:

$$\begin{aligned}(\sigma(u_x^j) - \sigma(u_x^{m,j}))\cdot u_{xt}^j &\le \sigma'(c^*)(u_x^j - u_x^{m,j})\cdot u_{xt}^j \\ &= (jm-t)\{-\sigma'(c^*)\}\frac{(u_x^{m,j} - u_x^{m,j-1})^2}{m^2} \\ &\le m\cdot\max_{y\in\mathbb{R}}\{-\sigma'(y)\}\frac{(u_x^{m,j} - u_x^{m,j-1})^2}{m^2} \\ &= mL(v_x^{m,j})^2 \qquad (4.5)\end{aligned}$$

for some c^* between u_x^j and $u_x^{m,j}$. Moreover, both the second and the fourth term of (4.4) are negative since $t - jm \le 0$. Now the following inequalities on the energy $E(t)$ are derived:

$$\begin{aligned}\frac{d}{dt}E(t) &\le (-1+mL)\int_0^1 |v_x^{m,j}|^2 dx \\ &\le -\frac{1}{2}\|v_x^{m,j}\|_{L^2}^2\end{aligned}$$

for all $t \in ((j-1)m, jm]$. ∎

Note that by taking an integral from $(j-1)m$ to jm on both sides of the above inequality, we get

$$E(jm) - E((j-1)m) \le -\frac{1}{2} m \|v_x^{m,j}\|_{L^2}^2,$$

and after taking a summation from $j = 1$ to $j = S$, the following estimate is established:

$$E(Sm) - E(0) \le -\frac{m}{2} \sum_{j=1}^{S} \|v_x^{m,j}\|_{L^2}^2.$$

Therefore,

$$\frac{m}{2} \sum_{j=1}^{S} \|v_x^{m,j}\|_{L^2}^2 \le E(0) - E(Sm) < E(0) < \epsilon \tag{4.6}$$

for all $S \in \mathbb{N}$. Next, we show the several estimates on the functions which will play a significant role for the proof of Theorem 3.1.

LEMMA 4.2. *The following a priori estimates hold:*

(a) $\sup_{j\in\mathbb{N}} \sup_{(j-1)m<t\le jm} \|p^j(\cdot,t)\|_{L^\infty} \le \eta$,

(b) $\sup_{j\in\mathbb{N}} \sup_{(j-1)m<t\le jm} \|\pi_a(\int_0^x u^j(y,t)dy)\|_{L^\infty} \le \sigma_0\eta$, *where* $\pi_a(f) := f - \int_0^1 f$,

(c) $\sup_{j\in\mathbb{N}} \sup_{(j-1)m<t\le jm} \|q^j(\cdot,t)\|_{L^\infty} \le \tilde{K}$,

(d) $\sup_{j\in\mathbb{N}} \sup_{(j-1)m<t\le jm} \|u^j(\cdot,t)\|_{L^\infty} \le \tilde{K}$,

(e) $\sup_{j\in\mathbb{N}} \sup_{(j-1)m<t\le jm} |\int_0^1 \sigma(u_x^j(x,t))dx| \le \sigma_0\eta$,

(f) $\sup_{j\in\mathbb{N}} \sup_{(j-1)m<t\le jm} \|v^j(\cdot,t)\|_{L^\infty} = \sup_{j\in\mathbb{N}} \sup_{(j-1)m<t\le jm} \|p_x^j(\cdot,t)\|_{L^\infty} \le \tilde{K}$

for some constants $\tilde{K},\ \eta > 0,\ \eta \ll 1$.

Proof. Since $\int_0^1 p^j(x,t)dx = 0$, $p^j(x',t) = 0$ for some x' in $(0,1)$. Hence, the following holds:

$$|p^j(x,t)| = \left| \int_{x'}^{x} p_x^j(y,t)dy \right| \le \left(\int_0^1 |p_x^j(y,t)|^2 dy \right)^{\frac{1}{2}}.$$

Since $\|p_x^j(\cdot,t)\|_{L^2} = \|v^j(\cdot,t)\|_{L^2} \le \sqrt{E(u^j,v^j)} \le \sqrt{E(u_0,v_0)} \le \sqrt{\epsilon}$, (a) is accomplished by choosing $\eta > 0$ such that $\eta > \max\{\sqrt{\epsilon}, 2\sqrt{\epsilon}/\sigma_0, C_5\sqrt{\epsilon}/\sigma_0\}$, where C_5 will be chosen later.

Similarly,

$$\begin{aligned}
\left\| \int_0^x u^j(y,t)dy \right\|_{L^\infty} &= \left\| \int_0^x \int_0^y (p^j(z,t) + q^j(z,t))dzdy \right\|_{L^\infty} \\
&\le \left\| \int_0^x (p^j(y,t) + q^j(y,t))dy \right\|_{L^2} \\
&= \|u^j(\cdot,t)\|_{L^2} \\
&\le \sqrt{\epsilon} \le \frac{\sigma_0\eta}{2},
\end{aligned}$$

which proves (b).

By using (4.1),

$$q_t^j = \frac{u_x^{m,j}(x) - u_x^{m,j-1}(x)}{m} - \int_0^x v_t^j + \int_0^1 \int_0^x v_t^j$$

$$= -\pi_a(\sigma(u_x^{m,j})) + \pi_a\left(\int_0^x u^{m,j}\right) \tag{4.7}$$

$$= -\sigma(p^{m,j} + q^{m,j}) + e_1^{m,j}, \tag{4.8}$$

where $e_1^{m,j} = \int_0^1 \sigma(u_x^{m,j}(x))dx + \pi_a(\int_0^x u^{m,j}(y)dy)$. From the hypotheses (H2) and (H3), $\sigma(z) \le W(z) + C_1$ for some $C_1 > 0, z \in \mathbb{R}$. This and estimate (b) imply

$$|e_1^{m,j}| \le \int_0^1 [|W(u_x^{m,j})| + C_1]dx + \sigma_0\eta \le \epsilon + C_1 + \sigma_0\eta < C_2 \tag{4.9}$$

for some $C_2 > 0$. Since $\|p^j\|_{L^\infty} < \eta$, from (4.8), $q_t^j < 0$ when $q^j \ge K_1$ and $q_t^j > 0$ when $q^j \le -K_1$ for some sufficiently large $K_1 > 0$. Hence, q^j is bounded. Let $\tilde{K} > \max\{\eta + K_1, K_2\}$, where K_2 will be chosen later. This completes the proof of (c).

Note that

$$\|u_x^j\|_{L^\infty} \le \|p^j\|_{L^\infty} + \|q^j\|_{L^\infty} \le \eta + K_1 < \tilde{K}. \tag{4.10}$$

Now, (d) clearly follows from (4.10).

Note that by (4.10),

$$|\sigma(u_x^j)| \le C_3 \tag{4.11}$$

for some $C_3 > 0$. Since q_t^j satisfies (4.8), (4.11) combined with (4.9) implies that q_t^j is uniformly bounded for all $j \in \mathbb{N}$ in L^∞ norm. Also, note that

$$|\sigma'(u_x^j)| \le C_4 \tag{4.12}$$

for some $C_4 > 0$. From the coercivity condition (H2) on W and σ and by estimate (4.10),

$$C_5 := \sup_{z \in [-\tilde{K}, \tilde{K}] \setminus \{z_-, z_+\}} \frac{|\sigma(z)|}{\sqrt{W(z)}}$$

is well defined and

$$\left|\int_0^1 \sigma(u_x^j(x,t))dx\right| \le \|\sigma(u_x^j)\|_{L^2} \le C_5\left(\int_0^1 |W(u_x^j)|\right)^{\frac{1}{2}} < C_5\sqrt{\epsilon} \le \sigma_0\eta,$$

which proves (e).

It will be shown next that $\|p_{xx}^j\|_{L^2}$ is uniformly bounded for all $j \in \mathbb{N}$ in order to prove (f).

Since

$$p_{xx}^j(x,t) = r^j(x,t) + s^j(x,t),$$

where

$$
\begin{aligned}
r^j(x,t) &:= \left(\frac{mj-t}{m}\right) p_t^{j-1}(x) + \left(\frac{t-m(j-1)}{m}\right) p_t^j(x),\\
s^j(x,t) &:= \left(\frac{mj-t}{m}\right) q_t^{j-1}(x) + \left(\frac{t-m(j-1)}{m}\right) q_t^j(x),
\end{aligned}
\tag{4.13}
$$

and $\|q_t^j\|_{L^\infty}$ is uniformly bounded for all $j \in \mathbb{N}$, one would only need to show that $\|p_t^j\|_{L^2}$ is uniformly bounded for all $j \in \mathbb{N}$. By (4.7) and the identity $u_{xt}^j = p_{xx}^{m,j}$, p_t^j satisfies the following equation:

$$
p_t^j = p_{xx}^{m,j} + \pi_a \left[\sigma(p^{m,j} + q^{m,j}) - \int_0^x \int_0^{x'} (p^{m,j} + q^{m,j})\right]. \tag{4.14}
$$

Let $f(p^{m,j}) := -q_t^j$. Note that

$$
\|f(p^{m,j})\|_{L^\infty} < M_1 \tag{4.15}
$$

for some $M_1 > 0$ since q_t^j is uniformly bounded. From (4.14),

$$
p^{m,j} - p^{m,j-1} = m\Delta p^{m,j} + mf(p^{m,j}),
$$

which implies

$$
\begin{aligned}
p^{m,j} &= \frac{p^{m,j-1}}{(1-m\Delta)} + \frac{m}{(1-m\Delta)} f(p^{m,j})\\
&= \frac{1}{(1-m\Delta)}\left[\frac{p^{m,j-2}}{(1-m\Delta)} + \frac{m}{(1-m\Delta)} f(p^{m,j-1})\right] + \frac{m}{(1-m\Delta)} f(p^{m,j})\\
&= \frac{p^{m,j-2}}{(1-m\Delta)^2} + m\left[\frac{f(p^{m,j-1})}{(1-m\Delta)^2} + \frac{f(p^{m,j})}{(1-m\Delta)}\right]\\
&\quad \cdots\\
&= \frac{p_0}{(1-m\Delta)^j} + m\left[\frac{f(p^{m,1})}{(1-m\Delta)^j} + \cdots + \frac{f(p^{m,j})}{(1-m\Delta)}\right].
\end{aligned}
$$

Therefore,

$$
p_t^j = \frac{p^{m,j} - p^{m,j-1}}{m} = \frac{\Delta p_0}{(1-m\Delta)^j} + m\sum_{k=1}^{j-1} \frac{\Delta f(p^{m,k})}{(1-m\Delta)^{j+1-k}} + \frac{f(p^{m,j})}{(1-m\Delta)}.
$$

By incorporating the inequality $\|\frac{\Delta}{(1-m\Delta)}\|_{L^2} \le 1$ and (4.15), the following inequalities occur:

$$
\begin{aligned}
\|p_t^j\|_{L^2} &\le \|\Delta p_0\|_{L^2} + m\sum_{k=1}^{j-1} \left\| \frac{1}{(1-m\Delta)^{j-k}} \cdot \frac{\Delta}{(1-m\Delta)} \cdot f(p^{m,k}) \right\|_{L^2} + \|f(p^{m,j})\|_{L^2}\\
&\le \|\Delta p_0\|_{L^2} + mM_1 \cdot \sum_{k=1}^{j-1} \frac{1}{(1-m\lambda_1)^{j-k}} + M_1\\
&\le \|\Delta p_0\|_{L^2} + \frac{M_1}{\lambda_1} \cdot \left[\frac{1}{(1-m\lambda_1)^{j-1}} - 1\right] + M_1\\
&\le \|\Delta p_0\|_{L^2} + \left(-\frac{1}{\lambda_1} + 1\right) \cdot M_1.
\end{aligned}
$$

Here, $\lambda_1 < 0$ is the largest eigenvalue of Δ. Therefore, $\|p_t^j\|_{L^2}$ is uniformly bounded for all $j \in \mathbb{N}$ and $\|p_x^j\|_{L^\infty} < K_2$ for some $K_2 > 0$. Proof of Lemma 4.2 is now complete. □

Remark. One can see from the proofs of Lemmas 4.1 and 4.2 that the energy decay and a priori estimates are independent of m for sufficiently small $m > 0$.

5. Equilibrium state as the limit of the solution as $j \to \infty$. We now introduce the function φ, which is called the phase function. This function will play an important role in proving the equilibrium state of the solution at the limit as $j \to \infty$. Fix $r > 0$, $r \ll 1$ such that for $\lambda \in [-r, r]$, the equation $\sigma(z) = \lambda$ has three different solutions $z_1(\lambda) < z_2(\lambda) < z_3(\lambda)$. Define

$$\varphi(z) = \begin{cases} i, & z \in \bigcup_{\lambda \in [-r,\, r]} z_i(\lambda), \quad i = 1, 2, 3, \\ \infty & \text{elsewhere.} \end{cases}$$

The next proposition states that the discretized solution $u^{m,j}$ converges in $W_0^{1,p}$ to an equilibrium state as j goes to infinity.

PROPOSITION 5.1. *Suppose* (H1)–(H3), (A1)–(A4) *hold. Then the solution* $(u^{m,j}, v^{m,j})$ *of* (2.1) *converges strongly in* $W_0^{1,p} \times L^2$ $(1 \le p < \infty)$ *to some equilibrium state* $(u_\star^m, 0) \in W_0^{1,\infty} \times L^2$ *as* $j \to \infty$.

Proof. The proof consists of several lemmas. The following lemma states that under some appropriate conditions on the elastic stress $\sigma(u_x^{m,j}(x)) - \int_0^x u^{m.j}$ and the phase function φ, the strain $u_x^{m,j}$ converges to an equilibrium state. We must be careful when choosing the pointwise representatives of $u_x^{m,j}$ since in the measure zero sets of $(0,1)$, we never know the behavior of the strain $u_x^{m,j}$. It is important to choose a good representative so that the limit state is continuous except for the finitely many points which are the zeros of the limit state.

LEMMA 5.2. *Assume there exists a full measure subset* $\tilde{\Omega} \in (0,1)$ *(measure of* $\tilde{\Omega}$ *is* 1*) and pointwise representatives* $\bar{w}^{m,j}$ *of* $u_x^{m,j}$ *such that*

(B1) $\sigma(\bar{w}^{m,j}(x)) - \int_0^x u^{m,j} =: \lambda_j^m(x) \to \lambda^m$ *as* $j \to \infty$ *for some* $\lambda^m \in (-r, r)$ *and all* $x \in \tilde{\Omega}$;

(B2) $\lim_{j\to\infty} \varphi(\bar{w}^{m,j}(x))$ *exists and is finite for all* $x \in \tilde{\Omega}$.

Then $\lim_{j\to\infty} \bar{w}^{m,j}(x) =: \bar{w}^m(x)$ *exists for all* $x \in \tilde{\Omega}$. *Moreover, the equivalence class* $\hat{w}^m$ *of* $\bar{w}^m$ *satisfies*

$$\|\hat{w}^m\|_{L_\infty} \le \tilde{K} \qquad \text{and} \qquad u_\star^m(x) := \int_0^x \hat{w}^m \quad \text{is in} \quad W_0^{1,\infty}.$$

Also,

$$\sigma((u_\star^m)_x(x)) - \int_0^x u_\star^m \equiv \lambda^m \ \ a.e., \qquad \varphi((u_\star^m)_x(x)) = \lim_{j\to\infty} \varphi(u_x^{m,j}(x)) \ \ a.e.$$

and

$$u^{m,j} \to u_\star^m \ \ in \ \ W_0^{1,p} \qquad (1 \le p < \infty).$$

Proof. Recall that $\sup_{j\in\mathbb{N}} \|u_x^j\|_{L^\infty} < \tilde{K}$ by (4.10). Define

$$\chi_i^{m,j}(x) := \begin{cases} 1, & x \in \tilde{\Omega} \ \text{ and } \ \varphi(\bar{w}^{m,j}(x)) = i \in \{1,2,3\}, \\ 0 & \text{else,} \end{cases}$$

$$\chi_\infty^{m,j}(x) := \begin{cases} 1, & x \in \tilde{\Omega} \text{ and } \varphi(\bar{w}^{m,j}(x)) = \infty, \\ 0 & \text{else.} \end{cases}$$

Let $x \in \tilde{\Omega}$. Since $\bar{w}^{m,j}(x) = z_i(\int_0^x u^{m,j} + \lambda_j^m(x))$ and $\chi_\infty^{m,j}(x) = 0$ if $\varphi(\bar{w}^{m,j}(x)) = i$, $i = 1, 2, 3$, the following equation holds:

$$\begin{aligned} &\bar{w}^{m,j}(x) - \bar{w}^{m,k}(x) \\ &\quad = \sum_{i=1}^{3} \left[\chi_i^{m,j}(x) \cdot z_i \left(\int_0^x u^{m,j} + \lambda_j^m(x) \right) - \chi_i^{m,k}(x) \cdot z_i \left(\int_0^x u^{m,k} + \lambda_k^m(x) \right) \right] \\ &\qquad + \chi_\infty^{m,j}(x) \cdot \bar{w}^{m,j}(x) - \chi_\infty^{m,k}(x) \cdot \bar{w}^{m,k}(x). \end{aligned} \tag{5.1}$$

Note that since $1 = \frac{d}{d\lambda^m}(\sigma(z_i(\lambda^m))) = \sigma'(z_i(\lambda^m)) \cdot z_i'(\lambda^m)$,

$$\begin{aligned} |z_i(a) - z_i(b)| &\le \sup_{x \in [-r,\, r]} |z_i'(x)| \cdot |a - b| \\ &= \sup_{x \in [-r,\, r]} \frac{1}{|\sigma'(z_i(x))|} \cdot |a - b| \le \frac{1}{\bar{M}} |a - b|, \end{aligned} \tag{5.2}$$

where $\bar{M} := \min_{z \in \sigma^{-1}([-r,r])} |\sigma'(z)|$. Let $\xi_{j,k}^m(x) = \int_0^x |u^{m,j} - u^{m,k}|$, $j, k \in \mathbb{N}$. Then

$$\begin{aligned} 0 \le \frac{d}{dx} \xi_{j,k}^m(x) &= \left| \int_0^x (u_x^{m,j} - u_x^{m,k}) \right| \\ &= \left| \int_0^x \left[\sum_{i=1}^{3} \left\{ \chi_i^{m,j}(x') \cdot z_i \left(\int_0^{x'} u^{m,j} + \lambda_j^m(x') \right) \right. \right. \right. \\ &\qquad \left. - \chi_i^{m,k}(x') \cdot z_i \left(\int_0^{x'} u^{m,k} + \lambda_k^m(x') \right) \right\} \\ &\qquad \left. + \chi_\infty^{m,j}(x') \cdot \bar{w}^{m,j}(x') - \chi_\infty^{m,k}(x') \cdot \bar{w}^{m,k}(x') \right] dx' \Bigg| \\ &= \left| \int_0^x \left[\sum_{i=1}^{3} \chi_i^{m,k}(x') \{ \bar{w}^{m,j}(x') - \bar{w}^{m,k}(x') \} + \sum_{i=1}^{3} \{ \chi_i^{m,j}(x') - \chi_i^{m,k}(x') \} \bar{w}^{m,j}(x') \right. \right. \\ &\qquad \left. \left. + \chi_\infty^{m,j}(x') \cdot \bar{w}^{m,j}(x') - \chi_\infty^{m,k}(x') \cdot \bar{w}^{m,k}(x') \right] dx' \right| \end{aligned} \tag{5.3}$$

holds for $j, k \in \mathbb{N}$. Note that

$$\int_0^x \left[\sum_{i=1}^{3} (\chi_i^{m,j} - \chi_i^{m,k}) \bar{w}^{m,j} \right] \le 2\tilde{K} |\{x \in (0,1) : \varphi(\bar{w}^{m,j}(x)) \ne \varphi(\bar{w}^{m,k}(x))\}|$$

and

$$\begin{aligned} \int_0^x \left(\chi_\infty^{m,j} \cdot \bar{w}^{m,j} - \chi_\infty^{m,k} \cdot \bar{w}^{m,k} \right) &\le \tilde{K} |\{x \in (0,1) : \varphi(\bar{w}^{m,j}(x)) \ne \varphi(\bar{w}^{m,k}(x))\}| \\ &\quad + 2\tilde{K} |\{x \in (0,1) : \varphi(\bar{w}^{m,j}(x)) = \varphi(\bar{w}^{m,k}(x)) = \infty\}|. \end{aligned}$$

Therefore, the last three terms in (5.3) are dominated by

$$\begin{aligned} &3\tilde{K} |\{x \in (0,1) : \varphi(\bar{w}^{m,j}(x)) \ne \varphi(\bar{w}^{m,k}(x))\}| \\ &+ 2\tilde{K} |\{x \in (0,1) : \varphi(\bar{w}^{m,j}(x)) = \varphi(\bar{w}^{m,k}(x)) = \infty\}| =: \delta_{j,k}^m. \end{aligned}$$

Let $x \in (0,1)$ be fixed. By assumption (B2), $\varphi(\bar{w}^{m,j}(x)) = \varphi(\bar{w}^{m,k}(x)) = i(x)$ for some $i(x) = 1,2,3$ if j,k are sufficiently large. Therefore, $\delta^m_{j,k} \to 0$ as $\min\{j,k\} \to \infty$. For each $i \in \{1,2,3\}$,

$$\int_0^x \chi_i^{m,k}(x')(\bar{w}^{m,j}(x') - \bar{w}^{m,k}(x'))dx' \le \int_{J_1(i)} |\bar{w}^{m,j}(x') - \bar{w}^{m,k}(x')|dx' + \int_{J_2(i)} |\bar{w}^{m,j}(x') - \bar{w}^{m,k}(x')|dx',$$

where

$$J_1(i) := \{x' \in (0,x) : \chi_i^{m,j}(x') = \chi_i^{m,k}(x') = 1\},$$

$$J_2(i) := \{x' \in (0,x) : \chi_i^{m,j}(x') = 0,\ \chi_i^{m,k}(x') = 1\}.$$

In the set $J_1(i)$,

$$\begin{aligned}|\bar{w}^{m,j}(x') - \bar{w}^{m,k}(x')| &= \left| z_i\left(\int_0^{x'} u^{m,j}(s)ds + \lambda_j^m(x')\right) - z_i\left(\int_0^{x'} u^{m,k}(s)ds + \lambda_k^m(x')\right)\right| \\ &\le \frac{1}{\bar{M}}\left[\int_0^{x'} |u^{m,j}(s) - u^{m,k}(s)|ds + |\lambda_j^m(x') - \lambda_k^m(x')|\right] \\ &= \frac{1}{\bar{M}}[\xi^m_{j,k}(x') + |\lambda_j^m(x') - \lambda_k^m(x')|].\end{aligned}$$

Note that $J_2(i) \subset \{x \in (0,1) : \varphi(\bar{w}^{m,j}(x)) \ne \varphi(\bar{w}^{m,k}(x))\}$, $i = 1,2,3$. Hence,

$$\begin{aligned}\frac{d}{dx}\xi^m_{j,k}(x) &\le 2\cdot\delta^m_{j,k} + \frac{1}{\bar{M}}\int_0^1 |\lambda_j^m(x') - \lambda_k^m(x')|dx' + \int_0^x \frac{1}{\bar{M}}\xi^m_{j,k}(x')dx' \\ &\le 2\cdot\delta^m_{j,k} + \frac{1}{\bar{M}}\|\lambda_j^m - \lambda_k^m\|_{L^1} + \frac{1}{\bar{M}}\xi^m_{j,k}(x) \\ &\le \epsilon^m_{j,k} + \frac{1}{\bar{M}}\xi^m_{j,k}(x),\end{aligned}$$

where $\epsilon^m_{j,k} := 2\cdot\delta^m_{j,k} + \frac{1}{\bar{M}}\|\lambda_j^m - \lambda_k^m\|_{L^1}$. By assumption (B1), $\epsilon^m_{j,k} \to 0$ as $\min\{j,k\} \to \infty$. By Gronwall's inequality,

$$\xi^m_{j,k}(x) \le \epsilon^m_{j,k}\cdot\bar{M}\cdot\left(\exp\left(\frac{1}{\bar{M}}x\right) - 1\right) \to 0 \ \text{ as } \ \min\{j,k\} \to \infty.$$

Therefore,

$$\left|\int_0^x u^{m,j} - \int_0^x u^{m,k}\right| \le \xi^m_{j,k}(x) \to 0$$

as $\min\{j,k\} \to \infty$. By combining this with assumption (B1), we get

$$\left|\left(\int_0^x u^{m,j} + \lambda_j^m(x)\right) - \left(\int_0^x u^{m,k} + \lambda_k^m(x)\right)\right| \to 0 \ \text{ a.e.}$$

as $\min\{j,k\} \to \infty$. By assumption (B2), $\chi_i^{m,j}(x) = \chi_i^{m,k}(x) = 1$ for some $i(x) = 1,2,3$, and $\chi_\infty^{m,j}(x) = \chi_\infty^{m,k}(x) = 0$ for sufficiently large $j,k \in \mathbb{N}$. This implies that

the right-hand side of (5.1) converges to zero, and thus $\bar{w}^{m,j}(x) - \bar{w}^{m,k}(x) \to 0$ for all $x \in \tilde{\Omega}$ as $\min\{j,k\} \to \infty$. Hence,

$$\lim_{j\to\infty} \int_0^x u^{m,j} =: U^m \quad \text{exists } \forall\ x \in [0,1],$$
$$\lim_{j\to\infty} \bar{w}^{m,j} =: \bar{w}^m \quad \text{exists } \forall\ x \in \tilde{\Omega}.$$

This implies that $u_x^{m,j}$ converges to the equivalence class $\hat{w}^m$ of $\bar{w}^m$ in L^1. Let $u_\star^m(x) := \int_0^x \hat{w}^m$. Since

$$u_\star^m(1) = \int_0^1 \hat{w}^m = \lim_{j\to\infty} \int_0^1 u_x^{m,j} = \lim_{j\to\infty} (u^{m,j}(1) - u^{m,j}(0)) = 0,$$

$u_\star^m$ satisfies the boundary conditions of (1.1). Moreover, since

$$u^{m,j}(x) = \int_0^x u_x^{m,j} \to \int_0^x \hat{w}^m = u_\star^m(x) \ \text{ in } \ C([0,1]),$$

$U^m = \int_0^x u_\star^m$. Hence,

$$u^{m,j} \to u_\star^m \ \text{ in } \ W^{1,p}, \qquad 1 \le p < \infty.$$

Since $u^{m,j},\ u_x^{m,j}$ are uniformly bounded, $u_\star^m \in W_0^{1,\infty}$. Therefore,

$$u_x^{m,j} \to (u_\star^m)_x \quad \text{boundedly a.e.}$$

Since $\sigma(u_x^{m,j}) - \int_0^x u^{m,j} \to \sigma((u_\star^m)_x) - \int_0^x u_\star^m$ boundedly a.e. by assumption (B1),

$$\lambda^m = \sigma((u_\star^m)_x) - \int_0^x u_\star^m \quad \text{a.e.} \tag{5.4}$$

Since $(u_\star^m)_x$ lies in one of the three intervals $\bigcup_{\lambda\in[-r,\,r]} z_i(\lambda),\ \ i \in \{1,2,3\}$ a.e., we can choose the nice pointwise representatives $\bar{w}^{m,j}$ of $u_x^{m,j}$ such that (5.4) holds for the set $(0,1)$ except for the finitely many points which are the limits $(x_\star)_i^m$ of finitely many zeros $x_i^m(j), i = 1,\dots,N$, of $u_x^{m,j}$ in (P4). Hence, we can conclude that $(u^{m,j}, v^{m,j})$ converges to an equilibrium state $(u_\star^m, 0)$ strongly in $W_0^{1,p} \times L^2$. ∎

In the lemmas to follow, we will show that under the low initial energy, the assumptions (B1) and (B2) are satisfied. Lemma 5.3 shows that the convergence of mean elastic stress $\int_0^1 (\sigma(u_x^{m,j}) - \int_0^x u^{m,j})dx$ implies the convergence of elastic stress $\sigma(u_x^{m,j}) - \int_0^x u^{m,j}$.

LEMMA 5.3. *Let $u^{m,j},\ j \in \mathbb{N}$, be a solution of* (2.1). *Assume that*

$$\lim_{j\to\infty} \int_0^1 \left(\sigma(u_x^{m,j}) - \int_0^x u^{m,j}\right) dx =: \lambda^m \quad \textit{exists.}$$

Then

$$\sigma(u_x^{m,j}) - \int_0^x u^{m,j} \to \lambda^m \ \ \textit{a.e. as} \ \ j \to \infty.$$

Proof. By (4.7), the sufficient condition to the conclusion is when q_t^j goes to zero a.e. as $j \to \infty$. Define the following modification of the energy functional $E(t)$:

$$\widetilde{E}(t) := \int_0^1 \left[W(u_x^j(x,t)) + \frac{1}{2}(u^j(x,t))^2 + p^j(x,t)s^j(x,t)\right] dx,$$

where $s^j(x,t)$ is the interpolation function defined in (4.13). Note that $\widetilde{E}(t)$ is uniformly bounded and, moreover, sufficiently small since the first two terms are the part of energy functional $E(t)$ and the third term is small since $p^j(x,t)$ is small enough by estimate (a) of Lemma 4.2 and $s^j(x,t)$ is uniformly bounded since q_t^j is uniformly bounded for all $j \in \mathbb{N}$. By (4.1) and the integration by parts,

$$
\begin{aligned}
\frac{d}{dt}\widetilde{E}(t) &= \int_0^1 [\,\sigma(u_x^j)\cdot u_{xt}^j + u^j \cdot u_t^j + p_t^j \cdot q_t^j + p_t^j(s^j - q_t^j) + p^j \cdot s_t^j]dx \\
&= \int_0^1 \Bigg[(\sigma(u_x^j) - \sigma(u_x^{m,j}))\cdot u_{xt}^j + (u^j - u^{m,j})\cdot u_t^j \\
&\qquad + \sigma(u_x^{m,j})\cdot u_{xt}^j + u^{m,j}\cdot u_t^j + u_{xt}^j \cdot q_t^j - (q_t^j)^2 \\
&\qquad + \ p_t^j\left(\left(\frac{mj-t}{m}\right)q_t^{m,j-1} + \left(\frac{t-m(j-1)-m}{m}\right)q_t^{m,j}\right) + p^j s_t^j\Bigg]\,dx \\
&\le mL\|v_x^{m,j}\|_{L^2}^2 + (t-mj)\|v^{m,j}\|_{L^2}^2 - \int_0^1 (q_t^j)^2 dx \\
&\qquad + \int_0^1 \Bigg[(v_{xx}^{m,j} - v_t^j)\cdot u_t^j + u_{xt}^j u_{xt}^j - u_{xt}^j p_t^j \\
&\qquad\qquad + \ p_t^j\left(\frac{t-mj}{m}\right)\cdot(q_t^{m,j} - q_t^{m,j-1}) + p^j s_t^j\Bigg]\,dx \qquad (5.5) \\
&\le mL\|v_x^{m,j}\|_{L^2}^2 - \int_0^1 (q_t^j)^2 dx \\
&\qquad + \int_0^1 [-u_{xt}^j v_x^{m,j} - u_t^j v_t^j + u_{xt}^j u_{xt}^j + u_t^j p_{xt}^j + p_t^j\cdot(t-mj)\cdot s_t^j + p^j s_t^j]dx \qquad (5.6) \\
&= mL\|v_x^{m,j}\|_{L^2}^2 - \int_0^1 (q_t^j)^2 dx + \int_0^1 s_t^j\left[(t-mj)\cdot\left(\frac{p^{m,j} - p^{m,j-1}}{m}\right)\right. \\
&\qquad \left. + \left(\frac{mj-t}{m}\right)p^{m,j-1} + \left(\frac{t-m(j-1)}{m}\right)p^{m,j}\right]dx \\
&= mL\|v_x^{m,j}\|_{L^2}^2 - \int_0^1 (q_t^j)^2 dx + 2\cdot\int_0^1 s_t^j\cdot p^j dx - \int_0^1 s_t^j \cdot p^{m,j} dx.
\end{aligned}
$$

The first term of (5.5) is followed from estimate (4.5). The first four terms of the integrand of the third term of (5.6) vanish because of the identities $u_{xt}^j = v_x^{m,j}$, $p_{xt}^j = v_t^j$. Since

$$
\begin{aligned}
\left|\int_0^1 s_t^j \cdot p^{m,j} dx\right| &\le \|p_{xx}^{m,j}\|_{L^2}\cdot\|s_t^j\|_{L^2} \\
&= \|v_x^{m,j}\|_{L^2}\cdot\left\|\frac{q_t^j - q_t^{j-1}}{m}\right\|_{L^2} \\
&\le \|v_x^{m,j}\|_{L^2}\cdot\left(\left\|\pi_a\left(\sigma'(c^{**})\cdot\frac{u_x^{m,j} - u_x^{m,j-1}}{m}\right)\right\|_{L^2} + \left\|\frac{u_x^{m,j} - u_x^{m,j-1}}{m}\right\|_{L^2}\right) \\
&\le M_2\cdot\|v_x^{m,j}\|_{L^2}^2,
\end{aligned}
$$

$$\left|\int_0^1 s_t^j \cdot p^j dx\right| \le \|v_x^j\|_{L^2} \cdot \|s_t^j\|_{L^2}$$
$$\le M_3 \cdot (\|v_x^{m,j-1}\|_{L^2} + \|v_x^{m,j}\|_{L^2}) \cdot \|v_x^{m,j}\|_{L^2},$$

and

$$\|v_x^{m,j-1}\|_{L^2} \cdot \|v_x^{m,j}\|_{L^2} \le M_4 \cdot (\|v_x^{m,j-1}\|_{L^2}^2 + \|v_x^{m,j}\|_{L^2}^2)$$

for some c^{**} between $u_x^{m,j}$ and $u_x^{m,j-1}$, M_2, M_3, and $M_4 > 0$, the following estimate on $\frac{d}{dt}\widetilde{E}(t)$ holds:

$$\frac{d}{dt}\widetilde{E}(t) \le mL\|v_x^{m,j}\|_{L^2}^2 - \int_0^1 (q_t^j)^2 dx + M_5 \cdot (\|v_x^{m,j-1}\|_{L^2}^2 + \|v_x^{m,j}\|_{L^2}^2) \tag{5.7}$$

for some $M_5 > 0$. By taking an integral from $(j-1)m$ to jm on both sides of (5.7), we get

$$\widetilde{E}(jm) - \widetilde{E}((j-1)m) \le -m\int_0^1 (q_t^j)^2 + (mL + M_5)m\|v_x^{m,j}\|_{L^2}^2 + mM_5\|v_x^{m,j-1}\|_{L^2}^2.$$

By taking the summation $j = 1, \dots, S$, we get the following estimate:

$$\widetilde{E}(Sm) - \widetilde{E}(0) \le -m\sum_{j=1}^{S}\int_0^1 (q_t^j)^2 + (mL + M_6)\sum_{j=1}^{S} m\|v_x^{m,j}\|_{L^2}^2 + mM_5\|(v_0)_x\|_{L^2}^2$$

for some $M_6 > 0$. By (4.6),

$$m\sum_{j=1}^{S}\int_0^1 (q_t^j)^2 \le \widetilde{E}(0) - \widetilde{E}(Sm) + 2(mL + M_6)\epsilon + M_5 m\|(v_0)_x\|_{L^2}^2$$
$$\le |\widetilde{E}(0)| + |\widetilde{E}(Sm)| + 2(mL + M_6)\epsilon + \epsilon_1$$
$$\le \delta$$

for some $\epsilon_1, \delta \ll 1$. Therefore, $m\sum_{j=1}^{\infty}\int_0^1 (q_t^j)^2 \le \delta$, and this implies $q_t^j \to 0$ a.e. as $j \to \infty$. ∎

The next lemma shows the convergence of the phase function under the assumptions of the low initial energy and the convergence of mean elastic stress.

LEMMA 5.4. *Let $u^{m,j}$, $j \in \mathbb{N}$, be the solution of* (2.1). *Then the assumption* (B2) *in Lemma* 5.2 *holds.*

Proof. By (b) and (e) of Lemma 4.2, mean elastic stress $\int_0^1 \left(\sigma(u_x^{m,j}) - \int_0^x u^{m,j}\right) dx$ is sufficiently small. Then by Lemma 5.3, $\limsup_{j\to\infty}\left|\sigma(u_x^{m,j}) - \int_0^x u^{m,j}\right|$ is sufficiently small a.e. Combining this and estimate (b) of Lemma 4.2, $\limsup_{j\to\infty}\left|\sigma(u_x^j)\right| \le \frac{2r}{3}$ a.e. This implies that for almost every x, there exists $J(x) \in \mathbb{N}$ such that

$$\{u_x^j(x,t) : j \ge J(x)\} \subseteq \sigma^{-1}([-r,\ r]) = \bigcup_{i=1}^{3}\bigcup_{\lambda\in[-r,\ r]} z_i(\lambda).$$

Since $\{u_x^j(x,t) : j \ge J(x)\}$ is connected, $u_x^j(x,t) \in \bigcup_{\lambda\in[-r,\ r]} z_i(\lambda)$ for all $j \ge J(x)$ and for some $i(x) = 1, 2$, or 3. This implies that $\lim_{j\to\infty}\varphi(u_x^j(x,t))$ exists and is finite a.e. Consequently, $\lim_{j\to\infty}\varphi(u_x^{m,j}(x))$ also exists and is finite a.e. ∎

The next lemma shows the convergence of mean elastic stress.

LEMMA 5.5. *Let $u^{m,j}$ be the solution of* (2.1). *Then*

$$\lim_{j\to\infty} \underbrace{\left[\int_0^1 \left(\sigma(u_x^{m,j}) - \int_0^x u^{m,j}\right) dx\right]}_{=:\, c(j)} \quad \textit{exists.}$$

Proof. Suppose this fails. Then there exists a subsequence $j_k \to \infty$ such that $c(j_k) \to \lambda^m$ and another subsequence $j_s \to \infty$ such that $c(j_s) \to \bar{\lambda}^m$ for some $\lambda^m,\ \bar{\lambda}^m \in [-\frac{2r}{3}, \frac{2r}{3}]$ and $\lambda^m < \bar{\lambda}^m$. Then by Lemma 5.3, $\sigma(u_x^{m,j_k}) - \int_0^x u^{m,j_k} \to \lambda^m$ a.e. as $j_k \to \infty$ and $\sigma(u_x^{m,j_s}) - \int_0^x u^{m,j_s} \to \bar{\lambda}^m$ a.e. as $j_s \to \infty$. Also by Lemma 5.4, $\lim_{j_k\to\infty} \varphi(u_x^{m,j_k})$, $\lim_{j_s\to\infty} \varphi(u_x^{m,j_s})$ exist and are finite, respectively. Hence, these satisfy the assumptions (B1), (B2) of Lemma 5.2, and therefore there exist $u^m,\ \bar{u}^m \in W^{1,\infty}$ such that

$$\sigma(u_x^m)_x - \int_0^x u^m \equiv \lambda^m \ \text{ a.e.}, \quad \sigma(\bar{u}_x^m)_x - \int_0^x \bar{u}^m \equiv \bar{\lambda}^m \ \text{ a.e.},$$

$$\varphi(u_x^m(x)) = \lim_{j_k\to\infty} \varphi(u_x^{m,j_k}(x)) \ \text{ a.e.}, \quad \text{and} \quad \varphi(\bar{u}_x^m(x)) = \lim_{j_s\to\infty} \varphi(u_x^{m,j_s}(x)) \ \text{ a.e.}$$

Note that $\varphi(u_x^m(x)) = \varphi(\bar{u}_x^m(x)) =: \varphi_\infty(x)$ since the limit of the phase function is independent of λ^m and $\bar{\lambda}^m$.

Consider the case where $\varphi_\infty(x) \in \{1, 3\}$ a.e. That is, the measure of the set $\Omega_u := \{x \in (0,1) : \varphi_\infty(x) = 2\}$ is zero. Now we introduce the following principle, whose proof was done in [16].

Comparison principle for weak solutions of the ordinary differential equation $\boldsymbol{\sigma(u_x)_x = u}$. Assume that $u,\ \bar{u} \in W^{1,\infty}$ satisfy

$$\sigma(u_x)_x - \int_0^x u \equiv \lambda \ \text{ a.e.}, \quad \sigma(\bar{u}_x)_x - \int_0^x \bar{u} \equiv \bar{\lambda} \ \text{ a.e.},$$

$\lambda < \bar{\lambda}$, $u(0) = \bar{u}(0) = 0$, $\sigma(u_x),\ \sigma(\bar{u}_x) \in [-r, r]$ a.e., $\varphi(u_x) = \varphi(\bar{u}_x)$ a.e., and $\varphi(u_x) \in \{1, 3\}$ a.e. Then $u(x) < \bar{u}(x)$ for all $x \in (0, 1]$.

Since u^m and $\bar{u}^m$ satisfy the assumptions of the above comparison principle, $u^m(1) < \bar{u}^m(1)$. This contradicts the boundary conditions of (1.1). In the case when the measure of Ω_u is not zero, contradiction arises from the following modified principle, which was also proven in [16].

Refined comparison principle for weak solutions of the ordinary differential equation $\boldsymbol{\sigma(u_x)_x = u}$. Under the same assumptions as the comparison principle, but with the condition $\varphi(u_x) \in \{1, 3\}$ a.e. replaced by $\int_0^1 W(u_x) < \epsilon$ and $|\Omega_u| \neq 0$, the inequality

$$u(1) < \bar{u}(1)$$

holds. □

Now Proposition 5.1 is complete. □

6. Dynamical behavior of the transition layers. If the set

$$\mathcal{L}_{\frac{\rho}{2}}(j) = \left\{x \in (0,1) : |u_x^j(x,t)| \le \frac{\rho}{2}\right\}$$

is monotonically decreasing to the finitely many isolated points as $j \to \infty$, we obtain the desired conclusion, since this is equivalent to the fact that the layers get steeper and eventually become discontinuous as j approaches infinity. However, the set $\mathcal{L}_{\frac{\rho}{2}}(j)$ is not decreasing as $j \to \infty$. We define the following set $\widetilde{\mathcal{L}}(j)$ instead and show that the set $\mathcal{L}_{\frac{\rho}{2}}(j)$ is contained in $\widetilde{\mathcal{L}}(j)$. We will then show that the set $\widetilde{\mathcal{L}}(j)$ is decreasing to the finitely many isolated points. Let $\eta \in (0, \frac{\rho}{4})$. Set $\rho_0 := \rho - \eta$. Define

$$\widetilde{\mathcal{L}}(j) := \{x \in (0,1) : |q^j(x,t)| \le \rho_0\}.$$

The following lemma states that the set of transition layers are always in the set $\widetilde{\mathcal{L}}(j)$ and furthermore in the set of initial transition layers $\mathcal{L}_\rho(0)$. This lemma plays an important role in showing the preservation of the number of transition layers.

LEMMA 6.1.

$$\mathcal{L}_{\frac{\rho}{2}}(j) \subseteq \widetilde{\mathcal{L}}(j) \subseteq \mathcal{L}_\rho(0) \quad \forall j \in \mathbb{N}.$$

Proof. If $x \in \mathcal{L}_{\frac{\rho}{2}}(j)$, then $|u_x^j(x,t)| \le \frac{\rho}{2}$. Therefore, by estimate (a) of Lemma 4.2,

$$|q^j(x,t)| = |u_x^j - p^j| \le \frac{\rho}{2} + \eta < \frac{\rho}{2} + \frac{\rho}{4} < \rho - \eta = \rho_0.$$

Now, $\widetilde{\mathcal{L}}(j) \subseteq \mathcal{L}_\rho(0)$ clearly follows. ∎

Next we show that the set $\widetilde{\mathcal{L}}(j)$ is exponentially decreasing to the finitely many isolated points.

LEMMA 6.2. *Assume $K > 4\tilde{K}$. Then for all $j \in \mathbb{N}$ and for some $C_0 > 0$,*

(i) $|q_x^j(x,t)| \ge C_0 e^{jm\sigma_0}|(q_0)_x|$ *if* $x \in \widetilde{\mathcal{L}}(j)$ *(exponential growth),*

(ii) $\widetilde{\mathcal{L}}(j+1) \subseteq \widetilde{\mathcal{L}}(j)$ *(monotonicity).*

Proof. We will show (i) by induction. Fix $j \in \mathbb{N}$ and fix $x \in \widetilde{\mathcal{L}}(j)$. Then $x \in \mathcal{L}_\rho(0)$ by Lemma 6.1. By hypothesis (A4), $|(u_0)_{xx}(x)| \ge K$. Suppose $(u_0)_{xx}(x) \ge K$. Since $(p_0)_x(x) < \tilde{K}$ by estimate (f) of Lemma 4.2, $(q_0)_x(x) = (u_0)_{xx}(x) - (p_0)_x(x) > 0$. By differentiating (4.7) with respect to x for $j = 1$, and by using the estimates (d), (f) of Lemma 4.2 and (4.12), we get the following estimate:

$$\begin{aligned} q_x^{m,1}(x) - q_x^{m,0}(x) &= \{-[\sigma(u_x^{m,1}(x))]_x + u^{m,1}(x)\}m \\ &= \{-\sigma'(u_x^{m,1}(x))(p_x^{m,1}(x) + q_x^{m,1}(x)) + u^{m,1}(x)\}m \\ &\ge -\sigma'(u_x^{m,1}(x))q_x^{m,1}(x)m - C_6 m \end{aligned}$$

for some $C_6 > 0$. Hence,

$$(1 + \sigma'(u_x^{m,1}(x))m)q_x^{m,1}(x) \ge q_x^{m,0}(x) - C_6 m.$$

Since m is sufficiently small and $q_x^{m,0} = (q_0)_x > 0$, $q_x^{m,1}$ is also positive. Therefore, the inequality

$$(1 - \sigma_0 m)q_x^{m,1}(x) \ge (1 + \sigma'(u_x^{m,1}(x))m)q_x^{m,1}(x) \ge q_x^{m,0}(x) - C_6 m$$

holds. Recall that $\sigma_0 = \min_{[-\rho,\rho]} |\sigma'|$. By induction, suppose $q_x^{m,j-1} > 0$. Then $q_x^{m,j} > 0$ and

$$(1-\sigma_0 m) q_x^{m,j}(x) \geq q_x^{m,j-1}(x) - C_6 m.$$

By iterating this, we obtain

$$\begin{aligned}
q_x^{m,j} &\geq \frac{1}{1-\sigma_0 m} \cdot q_x^{m,j-1} - C_6 m \cdot \frac{1}{1-\sigma_0 m} \\
&\geq \frac{1}{1-\sigma_0 m} \cdot \left[\frac{1}{1-\sigma_0 m} \cdot q_x^{m,j-2} - C_6 m \cdot \frac{1}{1-\sigma_0 m} \right] - C_6 m \cdot \frac{1}{1-\sigma_0 m} \\
&= \frac{1}{(1-\sigma_0 m)^2} \cdot q_x^{m,j-2} - C_6 m \left[\frac{1}{1-\sigma_0 m} + \frac{1}{(1-\sigma_0 m)^2} \right] \\
&\cdots \\
&= \frac{1}{(1-\sigma_0 m)^j} \cdot (q_0)_x - C_6 m \left[\frac{1}{1-\sigma_0 m} + \cdots + \frac{1}{(1-\sigma_0 m)^j} \right] \\
&= \frac{1}{(1-\sigma_0 m)^j} \cdot \left((q_0)_x - \frac{C_6}{\sigma_0} \right) + \frac{C_6}{\sigma_0}.
\end{aligned}$$

This implies

$$q_x^{m,j} \geq e^{jm\sigma_0} \cdot (q_0)_x.$$

Therefore, we can establish the exponential growth of q_x^j, that is,

$$|q_x^j(x,t)| \geq C_0 e^{jm\sigma_0} \cdot |(q_0)_x|$$

for some $C_0 > 0$. Similarly, we get the same conclusion for the case $(u_0)_{xx}(x) \leq -K$, and this proves (i) of Lemma 6.2.

Note that for $K > 4\tilde{K}$,

$$|q_x^j(x,t)| \geq C_0 e^{jm\sigma_0} \cdot |(q_0)_x| \geq C_0 e^{jm\sigma_0} (|(u_0)_{xx}| - \tilde{K}) \geq 3K_0 e^{jm\sigma_0}. \tag{6.1}$$

Here, $K_0 = \tilde{K} C_0$. If $q^j = \rho_0$, then $u_x^j = p^j + q^j = p^j + \rho_0 \geq -\eta + \rho_0 > 0$, and if $q^j = -\rho_0$, then $u_x^j = p^j - \rho_0 \leq \eta - \rho_0 < 0$, which implies $sign(u_x^j) = sign(q^j)$ at $|q^j| = \rho_0$. By using this and (4.7), and also by using estimates (b), (e) of Lemma 4.2, we have the estimate

$$\begin{aligned}
\frac{d}{dt}|q^j(x,t)| &= sign(q^j(x,t)) \cdot \left[\frac{q^{m,j}(x) - q^{m,j-1}(x)}{m} \right] \\
&= sign(u_x^j(x,t)) \cdot \left[\sigma(0) - \sigma(u_x^{m,j}(x)) + \int_0^1 \sigma(u_x^{m,j}) + \pi_a \left(\int_0^x u^{m,j} \right) \right] \\
&\geq -\sigma'(c') \cdot u_x^j \cdot sign(u_x^j) - \sigma'(c') \cdot (u_x^{m,j} - u_x^j) \cdot sign(u_x^j) - 2\sigma_0 \eta \\
&\geq \sigma_0 \cdot |u_x^j| - 2\sigma_0 \eta - \sigma'(c') \cdot sign(u_x^j) \cdot \frac{jm-t}{m} (u_x^{m,j} - u_x^{m,j-1}) \\
&\geq \sigma_0 \cdot (\rho - 4\eta) - \sigma'(c') \cdot sign(u_x^j) \cdot \frac{jm-t}{m} (u_x^{m,j} - u_x^{m,j-1}) \qquad (6.2)
\end{aligned}$$

at $|q^j| = \rho_0$ and for some c' between 0 and $u_x^{m,j}(x)$. Note that $|\frac{jm-t}{m}| < 1$. By estimate (a) of Lemma 4.2, $|p^{m,j} - p^{m,j-1}| \leq 2\eta \ll 1$. By (4.15), $|q^{m,j} - q^{m,j-1}| = m|q_t^j| \leq$

$mM_1 \ll 1$ when $m \ll 1$. Hence, $|u_x^{m,j} - u_x^{m,j-1}| \ll 1$, and this enables the second term of (6.2) to be small. Therefore,

$$\frac{d}{dt}|q^j(x,t)| \geq 0,$$

which implies $|q^j(x,t)| \leq |q^{j+1}(x,t)|$ for all $j \in \mathbb{N}$ when $|q^j(x,t)| = \rho_0$. By (i), q^j is strictly increasing or decreasing on $\widetilde{\mathcal{L}}(j)$, which implies (ii). □

From part (i) of Lemma 6.2, estimate (f) of Lemma 4.2, and the hypothesis (A4),

$$\begin{aligned} |u_{xx}^j(x,t)| &\geq |q_x^j(x,t)| - |p_x^j(x,t)| \\ &\geq C_0 e^{jm\sigma_0} \cdot |(q_0)_x| - \tilde{K} \\ &\geq C_0 e^{jm\sigma_0} \cdot (|(u_0)_{xx}| - 2\tilde{K}) \\ &\geq \frac{1}{2} K_0 e^{jm\sigma_0} \end{aligned} \tag{6.3}$$

if $x \in \mathcal{L}_{\frac{\rho}{2}}(j)$ and $K > 4\tilde{K}$.

From (6.3) and the fact that $\|u^{m,j}\|_{C^2} < \infty$ for all $j \in \mathbb{N}$, $\mathcal{L}_{\frac{\rho}{2}}(j)$ has a finite number of components $[a_i^m(j), b_i^m(j)]$, $0 < a_1^m(j) < b_1^m(j) < \cdots < a_N^m(j) < b_N^m(j) < 1$, in each of which $u_x^j(x,t)$ is strictly monotone and has exactly one zero $x_i^m(j)$. Also, $N(j) \geq 1$ since $u^j(0,t) = u^j(1,t) = 0$ for all $j \in \mathbb{N}$.

LEMMA 6.3. *$N(j) \equiv$ const. for all $j \in \mathbb{N}$.*

Proof. For all $j \in \mathbb{N}$, define

$$g^j(x,t) := u_x^j(x,t), \quad (j-1)m < t \leq jm.$$

Since $g^j, g_x^j \in C((0,1) \times ((j-1)m, jm])$ and at each zero (x_0,t_0) of g^j, $|g_x^j(x,t)| \geq \frac{K_0}{2} > 0$ by inequality (6.3), $\{g^j(x_0,t_0) | (x_0,t_0)$ is a zero of $g^j\}$ does not contain a critical value of $g^j(\cdot,t)$ for each t_0. By the implicit function theorem, the number of zeros of $g^j(\cdot,t)$ is independent of t for $(j-1)m < t \leq jm$ for all $j \in \mathbb{N}$. □

Similarly, by defining

$$g^j(x,t) := u_x^j(x,t) - \frac{\rho}{2} \quad \text{and} \quad g^j(x,t) := u_x^j(x,t) + \frac{\rho}{2},$$

the number of connected components of $\mathcal{L}_{\frac{\rho}{2}}(j)$ is independent of j. Now, the proof of (P1) and (P2) is complete.

From Lemma 6.1, $[a_i^m(j), b_i^m(j)] \subseteq [(a_0)_i, (b_0)_i]$, $i = 1, \ldots, N$. Moreover,

$$\begin{aligned} \rho &= |u_x^j(b_i^m(j),t) - u_x^j(a_i^m(j),t)| \\ &= \int_{a_i^m(j)}^{b_i^m(j)} |u_{xx}^j| dx \\ &\geq \frac{1}{2} K_0 e^{jm\sigma_0} \cdot |b_i^m(j) - a_i^m(j)|, \end{aligned}$$

which implies

$$|b_i^m(j) - a_i^m(j)| \leq \frac{2\rho}{K_0} \cdot e^{-jm\sigma_0} \quad \text{for all } i = 1, \ldots, N$$

for fixed j and $K > 4\tilde{K}$. This proves the last part of (P3). The rest of (P3) was already proved.

From (6.1) and from similar analysis as in the case $\mathcal{L}_{\frac{\rho}{2}}(j)$, $\widetilde{\mathcal{L}}(j)$ has a finite number of components $[\alpha_i^m(j), \beta_i^m(j)]$, $0 < \alpha_1^m(j) < \beta_1^m(j) < \cdots < \alpha_N^m(j) < \beta_N^m(j) < 1$. By Lemma 6.1, $x_i^m(j) \in [a_i^m(j), b_i^m(j)] \subseteq [\alpha_i^m(j), \beta_i^m(j)] \subseteq [a_i^0, b_i^0]$. By (ii) of Lemma 6.2, $[\alpha_i^m(j+1), \beta_i^m(j+1)] \subseteq [\alpha_i^m(j), \beta_i^m(j)]$. Therefore, the set of $[\alpha_i^m(j+1), \beta_i^m(j+1)]$ forms a nested family of intervals. Hence,

$$\begin{aligned} 2\rho > 2\rho_0 &= |q^j(\beta_i^m(j), t) - q^j(\alpha_i^m(j), t)| \\ &= \int_{\alpha_i^m(j)}^{\beta_i^m(j)} |q_x^j| dx \\ &\geq 3K_0 |\beta_i^m(j) - \alpha_i^m(j)| \cdot e^{jm\sigma_0}, \end{aligned}$$

which concludes the proof of (P4).

(P3) and (P4) automatically imply that $(u_\star^m)_x$ is discontinuous at every $(x_\star)_i^m$. It remains now to show that $(u_\star^m)_x$ is continuous on $(0,1)\backslash\{(x_\star)_1^m, \ldots, (x_\star)_N^m\}$. Since $u_\star^m$ is an equilibrium state, it satisfies the equation

$$\sigma((u_\star^m)_x(x)) = \int_0^x (u_\star^m) + \lambda^m$$

for some constant $\lambda^m > 0$. We know that the first term on the right-hand side of the above equation is small by estimate (b) of Lemma 4.2. Furthermore, λ^m is sufficiently small on $(0,1)\backslash\{(x_\star)_1^m, \ldots, (x_\star)_N^m\}$. Therefore, $(u_\star^m)_x$, the inverse image of σ, is continuous on those intervals, which proves (P5). Theorem 3.1 is finally complete.

Remark. The results of transition layer dynamics work for the discretized viscoelastic system without the elastic foundation term u, that is, for the system

$$\frac{1}{m^2}(u - 2u^{m,j-1} + u^{m,j-2}) - (\sigma(u_x))_x - \frac{1}{m}(u_x - u_x^{m,j-1})_x = 0.$$

The proof is similar to the proof for the system with the elastic foundation. Only the minor change of the proof of energy decay (Lemma 4.1), the proof of Lemma 5.3, and the estimate of $q^j(x,t)$ is needed.

7. Asymptotic behavior of the original system. In this section, we answer the following question: How do our results relate to the asymptotic behavior of the original system (1.1)?

We proved in section 5 that $u^{m,j}$ converges strongly in $W_0^{1,p}$ to a steady state $u_\star^m$ as $j \to \infty$ for fixed $m \ll 1$. Therefore, $u_\star^m$ satisfies

$$-(\sigma(u_x))_x + u = 0.$$

We will show next the existence of a weak limit of $u_\star^{m_k}$ in $W_0^{1,p}$ as $m_k \to 0$ for some sequence $m_k \ll 1$, $k \in \mathbb{N}$, in the following theorem.

THEOREM 7.1. *There is a sequence $m_k \ll 1$, $k \in \mathbb{N}$, and $m_k \to 0$ as $k \to \infty$ such that the steady state $u_\star^{m_k}$ in Proposition* 5.1 *converges in $W_0^{1,p}$ to a weak limit $u_\star$ as $k \to \infty$.*

Proof. The difficulty arises due to the nonlinearity of σ. However, by the fact that $(u_\star^m)_x$ is uniformly bounded by $\tilde{K} > 0$ and the coercivity condition for σ in (H2), the inequalities

$$\begin{aligned} \int_0^1 \sigma((u_\star^{m_k})_x) \cdot \zeta_x dx &\leq \hat{M} \int_0^1 (|(u_\star^{m_k})_x|^{p-1} + 1) \cdot \zeta_x dx \\ &\leq \hat{M} \int_0^1 (\tilde{K}^{p-1} + 1) \cdot \zeta_x dx \end{aligned}$$

hold for some $\hat{M} > 0$ and for any test function $\zeta \in C_0^\infty((0,1),\mathbb{R})$. The result follows from the dominated convergence theorem. ∎

Note that we can assume that σ is globally Lipschitz continuous since $u_x^{m,j}$ is uniformly bounded for all $j \in \mathbb{N}$ and for any $m \ll 1$. Then by [15, section 5.1], the weak solution of the system (1.1) is unique. Combining this with the results shown in [15, Theorem 4.1] and [16, Theorem 3.1] , the discretized solution $u^{m,j}$ converges in $W_0^{1,p}$ to a unique weak solution u of (1.1) as $m \to 0$, and u converges strongly in $W_0^{1,p}$ to a unique equilibrium state u_∞ as $t \to \infty$.

If the weak limit $u_\star$ is unique and is the same as u_∞, the same asymptotic behavior will hold for the system (1.1). However, we do not know the answer to this question and it remains as an open problem.

Acknowledgments. The author thanks Professor Zhengfang Zhou from Michigan State University for his valuable advice on this paper. The author also thanks the referees for their helpful comments and suggestions.

REFERENCES

[1] G. Andrews, *On the existence of solutions to the equation $u_{tt} = u_{xxt} + \sigma(u_x)_x$*, J. Differential Equations, 35 (1980), pp. 200–231.

[2] G. Andrews and J. M. Ball, *Asymptotic behavior and changes of phase in one-dimensional nonlinear viscoelasticity*, J. Differential Equations, 44 (1982), pp. 306–341.

[3] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Ration. Mech. Anal., 63 (1977), pp. 337–403.

[4] J. M. Ball, P. J. Holmes, R. D. James, R. L. Pego, and P. J. Swart, *On the dynamics of fine structure*, J. Nonlinear Sci., 1 (1991), pp. 17–70.

[5] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Ration. Mech. Anal., 100 (1987), pp. 13–52.

[6] J. M. Ball and R. D. James, *Proposed experimental tests of a theory of fine microstructure and the two-well problem*, Philos. Trans. Roy. Soc. London Ser. A, 338 (1992), pp. 389–450.

[7] H. Bellout and J. Nečas, *Existence of global weak solutions for a class of quasilinear hyperbolic integro-differential equations describing visco-elastic materials*, Math. Ann., 299 (1994), pp. 275–291.

[8] F. Bethuel, J. Coron, J. Ghidaglia, and A. Soyeur, *Heat flows and relaxed energies for harmonic maps*, in Nonlinear Diffusion Equations and Their Equilibrium States, Vol. 3, N. G. Lloyd, W. N. Ni, L. A. Peletier, and J. Serrin, eds., Progr. Nonlinear Differential Equations Appl. 7, Birkhäuser Boston, Boston, 1992, pp. 99–109.

[9] J. Clements, *Existence theorems for a quasilinear evolution equation*, SIAM J. Appl. Math., 26 (1974), pp. 745–752.

[10] S. Demoulini, *Young measure solutions for nonlinear evolutionary systems of mixed type*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 143–162.

[11] H. Engler, *Strong solutions for strongly damped quasilinear wave equations*, Contemp. Math., 64 (1987), pp. 219–237.

[12] J. L. Ericksen, *Equilibrium of bars*, J. Elasticity, 5 (1975), pp. 191–202.

[13] A. Friedman and J. Nečas, *Systems of nonlinear wave equations with nonlinear viscosity*, Pacific J. Math., 132 (1988), pp. 29–55.

[14] G. Friesecke, *A necessary and sufficient condition for nonattainment and formation of microstructure almost everywhere in scalar variational problems*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 437–471.

[15] G. Friesecke and G. Dolzmann, *Implicit time discretization and global existence for a quasilinear evolution equation with nonconvex energy*, SIAM J. Math. Anal., 28 (1997), pp. 363–380.

[16] G. Friesecke and J. B. McLeod, *Dynamics as a mechanism preventing the formation of finer and finer microstructure*, Arch. Ration. Mech. Anal., 133 (1996), pp. 199–247.

[17] J. M. Greenberg, *On the existence, uniqueness and stability of solutions of the equation $\rho_0 X_{tt} = E(X_x)X_{xx} + \lambda X_{xxt}$*, J. Math. Anal. Appl., 25 (1969), pp. 575–591.

[18] J. M. Greenberg, R. C. MacCamy, and V. J. Mizel, *On the existence, uniqueness and stability of solutions of the equation $\sigma'(u_x)u_{xx} + \lambda u_{xtx} = \rho u_{tt}$*, J. Math. Mech., 17 (1968),

pp. 707–728.
[19] K. Horihata and N. Kikuchi, *A construction of solutions satisfying a Cacciopoli inequality for non-linear parabolic equations associated to a variational functional of harmonic type*, Boll. Un. Mat. Ital. A (7), 3 (1989), pp. 199–207.
[20] H. Lim, *Time Discretization of Transition Layer Dynamics in Viscoelastic Systems*, Dissertation, Michigan State University, East Lansing, MI, 2001.
[21] M. Niezgodka and J. Sprekels, *Existence of solutions of a mathematical model of structural phase transitions in shape memory alloys*, Math. Methods Appl. Sci., 10 (1988), pp. 197–223.
[22] J. A. Nohel and R. L. Pego, *Nonlinear stability and asymptotic behavior of shearing motions of a non-Newtonian fluid*, SIAM J. Math. Anal., 24 (1993), pp. 911–942.
[23] J. A. Nohel, R. L. Pego, and A. E. Tzavaras, *Stability of discontinuous steady states in shearing motions of a non-Newtonian fluid*, Proc. Roy. Soc. Edinburgh Sect. A, 115 (1990), pp. 39–59.
[24] H. Pecher, *On global regular solution of third order partial differential equations*, J. Math. Anal. Appl., 73 (1980), pp. 278–299.
[25] R. L. Pego, *Phase transitions in one-dimensional nonlinear viscoelasticity: Admissibility and stability*, Arch. Ration. Mech. Anal., 97 (1987), pp. 353–394.
[26] M. Potier-Ferry, *On the mathematical foundations of elastic stability.* I, Arch. Ration. Mech. Anal., 78 (1982), pp. 55–72.
[27] P. Rybka, *Dynamical modeling of phase transitions by means of viscoelasticity in many dimensions*, Proc. Roy. Soc. Edinburgh Sect. A, 121 (1992), pp. 101–138.
[28] J. Sprekels and S. Zheng, *Global solutions to the equations of a Ginzburg-Landau theory for structural phase transitions in shape memory alloys*, Phys. D, 39 (1989), pp. 59–76.
[29] P. Swart, *The Dynamical Creation of Microstructure in Material Phase Transitions*, Dissertation, Cornell University, Ithaca, NY, 1991.
[30] P. J. Swart and P. J. Holmes, *Energy minimization and the formation of microstructure in dynamic anti-plane shear*, Arch. Ration. Mech. Anal., 121 (1992), pp. 37–85.
[31] A. Vainchtein, *Dynamics of phase transitions and hysteresis in a viscoelastic Ericksen's bar on an elastic foundation*, J. Elasticity, 57 (1999), pp. 243–280.
[32] A. Vainchtein, T. Healey, P. Rosakis, and L. Truskinovsky, *The role of the spinodal region in one-dimensional martensitic phase transitions*, Phys. D, 115 (1998), pp. 29–48.
[33] A. Vainchtein and P. Rosakis, *Hysteresis and stick-slip motion of phase boundaries in dynamic models of phase transitions*, J. Nonlinear Sci., 9 (1999), pp. 697–719.

SIAM J. MATH. ANAL.
Vol. 34, No. 3, pp. 595–610

CONVERGENCE OF VISCOSITY SOLUTIONS FOR ISOTHERMAL GAS DYNAMICS*

FEIMIN HUANG† AND ZHEN WANG‡

Abstract. We study the hyperbolic system of Euler equations for an isothermal, compressible fluid. The *strong convergence theorem* of approximate solutions is proved by the theory of compensated compactness. The existence of a weak entropy solution to Cauchy problems with large L^∞ initial data which may include a vacuum is also obtained. We note that we establish the commutation relations not only for the *weak* entropies but also for the *strong* ones by using the *analytic extension theorem*.

Key words. isothermal gas dynamics, compensated compactness, analytic extension theorem

AMS subject classifications. 35D05, 35L60, 35L65

PII. S0036141002405819

1. Introduction.

The one-dimensional Euler equations of compressible fluid read

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + p(\rho))_x = 0, \end{cases} \tag{1.1}$$

where the unknown variable $\rho \geq 0$ denotes the density of the mass, u the velocity. Usually, it is convenient to study (1.1) by a new variable, the momentum $m = \rho u$. Thus (1.1) becomes

$$\begin{cases} \rho_t + m_x = 0, \\ m_t + \left(\dfrac{m^2}{\rho} + p(\rho) \right)_x = 0. \end{cases} \tag{1.2}$$

For polytropic perfect gas, $p = p_0 \rho^\gamma$ with γ the adiabatic exponent. Equation (1.1) is called isentropic gas dynamics for $\gamma > 1$, while it is called isothermal gas dynamics for $\gamma = 1$. Without loss of generality, p_0 is normalized to be 1 here.

One of the main difficulties for the mathematical analysis of (1.1) is the singularity at the vacuum $\rho = 0$. It is noted that the term ρu^2 is only Lipschitz continuous for $\gamma > 1$ near the vacuum, while it is not even Lipschitz continuous for $\gamma = 1$ due to the infiniteness of the velocity u. This shows that the isothermal case is completely different from the isentropic one. Another difficulty is the development of a shock wave in solutions of (1.1) no matter how smooth the initial data are.

*Received by the editors April 18, 2002; accepted for publication (in revised form) May 13, 2002; published electronically December 13, 2002.
http://www.siam.org/journals/sima/34-3/40581.html

†Institute of Applied Mathematics, AMSS, Academia Sinica, Beijing 100080, People's Republic of China, and Department of Mathematics, Graduate School of Science, Osaka University, Osaka 560-0043, Japan (fhuang@math.sci.osaka-u.ac.jp). The first draft was completed while this author visited S.I.S.S.A., Via Beirut n.2-4, 34010, Trieste, Italy. He was supported in part by the JSPS Research Fellowship for foreign researchers and Grant-in-Aid P-00269 for JSPS from the ministry of Education, Science, Sports and Culture of Japan.

‡Wuhan Institute of Physics and Mathematics, The Chinese Academy of Sciences, P.O. Box 71010, Wuhan 430071, China (mazwang@cityu.edu.hk).

Away from the vacuum, the first result on the existence of BV solutions with large initial data in the BV space was obtained in Nishida [18] by using the Glimm scheme [11] for $\gamma = 1$. Poupaud, Rascle, and Vila [20] made a great simplification and improved the results of [18] to the isothermal Euler–Poisson system. For the case of $\gamma > 1$, the existence of a BV solution was established in Nishida and Smoller [19] for large total variation with small $\gamma - 1$.

When a vacuum occurs, the first global existence for (1.1) with large initial data in L^∞ was established in Diperna [10] for $\gamma = 1 + \frac{2}{2n+1}, n \geq 2$, by the theory of compensated compactness. For the interval $(1, \frac{5}{3}]$, the existence was solved by Ding, Chen, and Luo [6, 7] and Chen [2]. Lions, Perthame, and Tadmor [14] and Lions, Perthame, and Souganidis [15] treated this problem for $\gamma > \frac{5}{3}$. More recently, Chen and Le Floch [4] studied the existence problem for general pressure, where p acts like γ-law $(1 < \gamma < 3)$ near the vacuum $\rho = 0$. The approach of [4] further simplified the proofs for the case of $\gamma > 1$.

The purpose of the present paper is to prove the existence of a global weak L^∞ solution for isothermal gas dynamics with large initial data by the theory of compensated compactness. Our initial data are

$$(\rho(0,x), m(0,x)) = (\rho_0(x), m_0(x)) \in L^\infty, \tag{1.3}$$

where the vacuum may occur. It is known that all effective approaches for isentropic gas dynamics are based on the subtle analysis for the Euler–Poisson–Darboux (EPD) equation, while the entropy equation of $\gamma = 1$ is not governed by the EPD equation. Therefore all previous approaches [2, 4, 6, 7, 10, 14, 15] fail here. We shall use a new approach to achieve our goal. We establish the commutation relations for some strong entropies even though we do not know whether these strong waves satisfy the H^{-1} compact condition or not. It should be noted here that the strong entropy is useless for isentropic gas dynamics (see Lions, Perthame, and Tadmor [14]).

In addition, we note that the vacuum may disappear in the solution to the Riemann problem for isothermal flow. It is well known that the centered rarefaction wave may disappear in the solution of the Riemann problem for the scalar conservation laws, but initial value problems are still investigated when the initial data includes the centered rarefaction wave for the scalar equation. Similarly, the Cauchy problem for isothermal gas dynamics with the vacuum initial data is also important (see [1]) and has been long standing for many years. According to our experience with the Riemann problem, we conjecture that the solution we get here for the Cauchy problem may not contain the vacuum when $t > 0$. However, it seems that it is not easy to prove the above conjecture.

Now we recall the definition of the weak entropy solution.

DEFINITION 1.1. *$(\rho, m)(x,t) \in L^\infty(R^2_+)$ is called a weak entropy solution of* (1.2) *if it holds, for any test function $\phi \in C_0^\infty(R^2_+)$, that*

$$\begin{cases} \displaystyle\iint_{t>0} (\rho\phi_t + m\phi_x)\, dxdt + \int_R \rho_0(x)\phi(x,0)\, dx = 0, \\ \displaystyle\iint_{t>0} m\phi_t + \left(\frac{m^2}{\rho} + p(\rho)\right)\phi_x\, dxdt + \int_R m_0(x)\phi(x,0)\, dx = 0, \end{cases} \tag{1.4}$$

and for any weak entropy pair $(\eta, q)(\rho, m)$ with convex $\eta(\rho, m)$,

$$\eta(\rho, m)_t + q(\rho, m)_x \leq 0 \tag{1.5}$$

holds in the sense of distributions. Here the entropy pair (η, q) *is determined by the additional conservation law*

$$\eta(\rho, m)_t + q(\rho, m)_x = 0 \tag{1.6}$$

for any smooth solution of (1.2), *and* weak *entropy is an entropy that vanishes at the vacuum.*

Our main results follow.

THEOREM 1.2 (existence theorem). *Let* $\gamma = 1$, *and assume that the initial data satisfy*

$$0 \le \rho_0(x) \le M, \quad |m_0(x)| \le \rho_0(x)(M + |\log \rho_0(x)|) \quad a.e.; \tag{1.7}$$

then there exists a global weak entropy solution of (1.2), (1.3) *satisfying*

$$0 \le \rho(x,t) \le C, \quad |m(x,t)| \le \rho(x,t)(C + |\log \rho(x,t)|) \quad a.e., \tag{1.8}$$

where C *only depends on* M.

Remark 1.3. It is known that Nishida [18] had established the existence of global BV solutions for the isothermal gas dynamics if the initial data $(\rho_0, u_0) \in BV(R)$ and $\rho_0 > c > 0$. But here our initial data is more rough, i.e., $(\rho_0, u_0) \in L^\infty(R)$. Furthermore, our initial data may include the vacuum.

THEOREM 1.4 (compactness framework). *Let* $\gamma = 1$ *and let* $(\rho^\varepsilon, m^\varepsilon)$ *be a sequence of approximate solutions of* (1.2) *satisfying* (1.8) *uniformly in* ε. *Assume that*

$$\partial_t \eta(\rho^\varepsilon, m^\varepsilon) + \partial_x q(\rho^\varepsilon, m^\varepsilon) \quad \text{is compact in} \quad H^{-1}_{loc} \tag{1.9}$$

holds for some (not all) weak entropies (η, q) *with*

$$\eta = \rho^{\frac{1}{1-\xi^2}} e^{\frac{\xi}{1-\xi^2} u}, \quad q = (u + \xi)\eta, \quad \xi \in (-1, 1); \tag{1.10}$$

then there exists a function $(\rho(x,t), m(x,t))$ *satisfying* (1.8) *such that, extracting a subsequence if necessary,*

$$(\rho^\varepsilon(x,t), m^\varepsilon(x,t)) \to (\rho(x,t), m(x,t)) \quad \text{in } L^p_{loc}(R^2_+) \tag{1.11}$$

for all $p \in [1, +\infty)$.

Remark 1.5. Theorem 1.4 is useful for studying the existence of a global weak L^∞ solution of (1.1) with source term for $\gamma = 1$; for instance, the compressible Euler equations with damping, the Euler–Poisson system, etc.

Before we explain our ideas, it is worthwhile to briefly recall the theory of compensated compactness. To prove the existence, one usually constructs a sequence of approximate solutions $(\rho^\varepsilon, m^\varepsilon)$ by using viscosity perturbation or a finite difference scheme, then extracts a strong subsequence of $(\rho^\varepsilon, m^\varepsilon)$, if necessary, to get the desired results. However, it is very difficult to get a strong convergent subsequence for (1.2). Usually it is easy to obtain the uniform boundness estimates for $(\rho^\varepsilon, m^\varepsilon)$, which indicates that extracting a weak convergent subsequence is available. It is well known that weak convergence alone is not sufficient for implying the existence of a weak solution due to the nonlinearity of (1.2). So some information on the derivative of $(\rho^\varepsilon, m^\varepsilon)$ is needed. Tartar [22] first applied the Young measure to introduce the commutation relations

$$\langle \nu_{x,t}, q_1\eta_2 - q_2\eta_1 \rangle = \langle \nu_{x,t}, q_1 \rangle \langle \nu_{x,t}, \eta_2 \rangle - \langle \nu_{x,t}, q_2 \rangle \langle \nu_{x,t}, \eta_1 \rangle, \tag{1.12}$$

with any two entropy pairs $(\eta_i, q_i), i = 1, 2$, for almost every (x, t) if

$$\eta_i(\rho^\varepsilon, m^\varepsilon) + q_i(\rho^\varepsilon, m^\varepsilon) \tag{1.13}$$

lie in a compact subset of H^{-1}_{loc} as ε vanishes. If the Young measure satisfying (1.12) reduces to a point mass for almost every (x, t), then the weak convergence becomes strong, and the existence of a weak solution is established. Therefore the H^{-1} compact condition is essential to the theory of compensated compactness.

For strictly hyperbolic systems with smooth flux, the H^{-1} compact condition is easy due to the uniform boundness of approximate solutions and Murat's lemma [17], provided that the system has a strictly convex entropy. However, for isentropic gas dynamics, not all entropy pairs can be applied to Tartar commutation relations (1.12), since only weak entropy pairs are known to satisfy the H^{-1} compact condition. As pointed out by Lions, Perthame, and Tadmor [14], strong entropies are useless for the isentropic case. Fortunately, since all weak entropies obey famous EPD equation, people (see [2, 6, 7, 10, 14, 15]) are able to imply that the Young measure is either a single point or a subset of the vacuum by careful entropy analysis for the EPD equation, and then prove the existence of a weak solution for the isentropic case. More precisely, in the proof of [2, 6, 7, 10], the heart of the matter is to construct the special weak entropies and apply them to the commutation relations. This is possible because (1.12) represents an imbalance of regularity: the operator on the left is more regular than the one on the right due to cancellation. The novel idea of applying the technique of fractional derivatives was introduced in [6, 7, 8]. A new analysis of (1.12) was proposed by Lions, Perthame, and Tadmor [14] and Lions, Perthame, and Souganidis [15] for $\gamma > 1$. Motivated by a kinetic formulation of (1.1), they made the crucial observation that the use of special weak entropies could be bypassed and (1.12) be directly expressed with the entropy kernel of EPD equation. We refer to [2, 6, 7, 10, 14, 15] for details. Even though only weak entropy pairs are used in the case $\gamma > 1$, Diperna [10] conjectured that it may be possible to establish the commutation relations for all entropy pairs, weak and strong ones. If it is true, the proof is quite simple (see [10]).

In general, to exploit the classical theory of compensated compactness, the following steps are necessary

(1) *to construct a sequence of approximate solutions and obtain the uniform boundness of approximate solutions;*

(2) *to establish the H^{-1} compact condition for infinite entropy pairs;*

(3) *to apply the* div-curl *lemma into all entropy pairs satisfying* (2) *to establish the commutation relations;*

(4) *to apply the commutation relations to reduce Young measure to a point mass for almost every (x, t).*

Similar to the isentropic case, it is also difficult for $\gamma = 1$ to prove the H^{-1} compact condition for strong entropy pairs (step 2). In fact, we do not know whether the strong entropy pairs satisfy the H^{-1} compact condition or not. Compared with the isentropic case, the main difficulties for the isothermal flow arise on the following two aspects: the infiniteness of eigenvalues due to the presence of a vacuum, and the fact that the entropy equation is not of EPD type. Among them, the second one is essential. Thus all approaches of [2, 6, 7, 10, 14, 15] fail here.

Since only weak entropies are known to satisfy H^{-1} compact condition for $\gamma = 1$, it seems that the strong entropies are useless, as in the isentropic case. However, the use of strong entropies is the key point of our proofs. The main novelty of this paper

is that we establish the commutation relations not only for the *weak* entropies but also for the *strong* ones by using the *analytic extension theorem* even though we do not know whether or not strong entropies satisfy H^{-1} compact conditions.

To achieve our goal, we first choose a special formula of entropies parameterized by a complex variable ξ. The formula includes both weak and strong entropies determined by the value of ξ. Then we prove that there exists a segment such that for any ξ belonging to the segment, the entropy pair is of weak type and satisfies the H^{-1} compact condition. Therefore the commutation relations are established for some weak entropies in this segment. It is observed that the two sides of (1.12) are regular for ξ. In fact, they are analytic functions with respect to ξ. So the commutation relations exactly hold for the whole complex space except two points $(-1, 0)$ and $(1, 0)$ due to the analytic extension theorem. It is noted that the entropies are strong if $|\xi| > 1$ (see (2.9)). Therefore Diperna's conjecture [10] is partially verified for isothermal flow; i.e., the commutation relations hold for some weak and strong entropy pairs. We note that the H^{-1} compact condition for strong entropies (step 2) can be bypassed. Since both weak and strong entropy pairs are applied to (1.12), we establish a strong convergence theorem of approximate solutions and prove the existence of a weak entropy solution for isothermal gas dynamics. Finally, it is worthwhile to point out that our approximate solutions are constructed by adding the viscosity perturbation to (1.1) due to the infiniteness of eigenvalues. It is observed that the eigenvalues $\lambda_1 = u - 1$ and $\lambda_2 = u + 1$ increase with the speed of $|\ln \rho|$. This indicates the possibility of constructing approximate solutions by numerical scheme. We will discuss this in the future.

This paper is organized as follows: In section 2, we give a formula of entropy pairs, parameterized by a complex variable ξ. In section 3, we study the viscosity solutions $(\rho^\varepsilon, m^\varepsilon)$ and prove that the weak entropy fields lie in a compact subset of H^{-1}_{loc} when $\xi \in (-1, 1)$ as ε vanishes. In section 4, we prove a strong convergence theorem of approximate solutions and obtain the existence of a weak solution for isothermal gas dynamics.

2. Entropy waves. This section is devoted to the entropy for isothermal flow. We recall that (η, q) is an entropy-flux pair if for any smooth solutions of (1.1), it satisfies an additional equation,

$$\eta_t(\rho, u) + q_x(\rho, u) = 0. \tag{2.1}$$

By definition, weak entropy is an entropy η that vanishes at the vacuum.

Let $\gamma = 1$. Equation (2.1) yields $\nabla\eta\nabla f = \nabla q$ with the flux $f = (\rho u, \rho u^2 + \rho)^T$, i.e.,

$$q_\rho = u\eta_\rho + \frac{1}{\rho}\eta_u, \quad q_u = \rho\eta_\rho + u\eta_u, \tag{2.2}$$

which indicates

$$\eta_{\rho\rho} = \frac{1}{\rho^2}\eta_{uu}. \tag{2.3}$$

We choose the form $\eta = h(\rho)e^{ku}$; then (2.3) implies

$$h'' - \frac{k^2}{\rho^2}h = 0. \tag{2.4}$$

Thus, we have $h(\rho)=\rho^m$ with $m(m-1)=k^2$. Now we consider the parameter k in the complex space. Let $k=\frac{\xi}{1-\xi^2}$, $\xi\in\mathbf{C}$; then $m=\frac{1}{1-\xi^2}$. Therefore we have the following formula of entropy pairs:

$$\begin{aligned}\eta&=\rho^{\frac{1}{1-\xi^2}}e^{\frac{\xi}{1-\xi^2}u},\\ q&=(u+\xi)\eta.\end{aligned}\tag{2.5}$$

We note that these entropies are analytic functions with respect to ξ. It is easy to see that the points $(-1,0)$ and $(1,0)$ are singular for (η,q).

On the other hand, it is convenient to introduce a coordinate system of Riemann invariants (w,z) with

$$\nabla w\cdot r_1=0,\quad \nabla z\cdot r_2=0,\tag{2.6}$$

where $r_1=(1,u-1)^T$, $r_2=(1,u+1)^T$ are the right eigenvectors of the Jacobian matrix of f: $\nabla f r_i=\lambda_i r_i$, $i=1,2$. In the setting of isothermal flow, the Riemann invariants read

$$w=\rho e^u,\qquad z=\rho e^{-u}.\tag{2.7}$$

Thus we rewrite (2.5) as

$$\begin{aligned}\eta&=w^{\frac{1}{2(1-\xi)}}z^{\frac{1}{2(1+\xi)}},\\ q&=(u+\xi)\eta.\end{aligned}\tag{2.8}$$

It is observed that the formula (2.8) includes two kinds of entropies determined by the new complex variable ξ. By definition, η is a weak entropy if and only if the following hold:

$$\mathrm{Re}\frac{1}{2(1-\xi)}>0,\qquad \mathrm{Re}\frac{1}{2(1+\xi)}>0,\tag{2.9}$$

i.e., $\xi\in\Omega_w=\{\xi\in\mathbf{C};-1<\mathrm{Re}\xi<1\}$.

In fact, we can get more information from (2.8). If we consider ξ in the real space, the entropies defined in (2.8) form a fundamental set of weak entropies for isothermal flow. In other words, we have the following.

LEMMA 2.1. *Let $\gamma=1$; then for any $\varphi(\xi)\in C_0^\infty(-1,1)$,*

$$\begin{aligned}\eta&=\int_{-1}^{1}\varphi(\xi)w^{\frac{1}{2(1-\xi)}}z^{\frac{1}{2(1+\xi)}}\,d\xi,\\ q&=\int_{-1}^{1}\varphi(\xi)(u+\xi)w^{\frac{1}{2(1-\xi)}}z^{\frac{1}{2(1+\xi)}}\,d\xi\end{aligned}\tag{2.10}$$

is a weak entropy-flux pair of (1.1).

3. Viscosity solutions. We consider the viscous perturbation of the isothermal flow,

$$\begin{cases}\rho^\varepsilon_t+m^\varepsilon_x=\varepsilon\rho^\varepsilon_{xx},\\ m^\varepsilon_t+\left(\dfrac{(m^\varepsilon)^2}{\rho^\varepsilon}+\rho^\varepsilon\right)_x=\varepsilon m^\varepsilon_{xx},\end{cases}\tag{3.1}$$

with initial data

$$(\rho^\varepsilon, m^\varepsilon)|_{t=0} = (\rho_0^\varepsilon(x), m_0^\varepsilon(x)), \tag{3.2}$$

where $(\rho_0^\varepsilon(x), m_0^\varepsilon(x))$ satisfy

$$\varepsilon \le \rho_0^\varepsilon(x) \le M, \qquad |m_0^\varepsilon(x)| \le M. \tag{3.3}$$

It is easy to see that (3.3) holds if $\rho_0^\varepsilon(x)$ is given by smoothing out $\rho_0(x)$ with a standard mollifier and adding ε.

In terms of the theory of the positive invariant region in Chueh, Conley, and Smoller [5, 21], it is easy to see that $\{(w,z);\ w \le \text{const},\ \ z \le \text{const}\}$ is the invariant region of (3.1), which indicates that the Riemann invariants w^ε, z^ε are uniformly bounded in L^∞. This implies that $(\rho^\varepsilon, m^\varepsilon)$ are also uniformly bounded in L^∞. It is noted that there always exists a local smooth solution for (3.1) due to Diperna [10]. In order to prove the existence of a smooth solution for (3.1) and (3.2), it is also important to obtain an a priori estimate of the lower bound for the density ρ^ε. Diperna first gave the lower bound by his Lemma 4.1 (see [10]), even though this lemma was stated in an incorrect way. Chen [3] fixed this lemma. On the other hand, Lu [16] also studied the lower bound for general pressure $p(\rho)$ by maximum principle in which the restriction of initial data on the infinity was not needed. Thus the uniform L^∞ estimates and the lower bound of ρ^ε adding the local existence theorem gives the following global existence result.

LEMMA 3.1. *If the initial data satisfy the condition* (3.3), *then for any fixed* $\varepsilon > 0$, *there exists a smooth solution for the Cauchy problem* (3.1), (3.2) *in* $R_T = R \times [0,T]$ *(for arbitrary* T*) which satisfies*

$$0 < c(\varepsilon,t) \le \rho^\varepsilon(x,t) \le C, \qquad |m^\varepsilon(x,t)| \le \rho^\varepsilon(x,t)(C + |\log \rho^\varepsilon(x,t)|), \tag{3.4}$$

where $c(\varepsilon,t)$ *is an appropriate function and* C *depends only on* M.

In order to apply the theory of compensated compactness, it is necessary to prove the divergence of weak entropy-flux pair is in a compact subset of H^{-1} as ε vanishes.

Take the form (η, q) as in (2.8), and let $\xi \in (-1,1)$; we compute

$$\begin{aligned}
\eta_{\rho\rho} &= \frac{\xi^2}{(1-\xi^2)^2}(1 - 2\xi u + u^2)\rho^{\frac{\xi^2}{1-\xi^2}-1} e^{\frac{\xi}{1-\xi^2}u} > 0,\\
\eta_{\rho m} &= \frac{\xi^2}{(1-\xi^2)^2}(\xi - u)\rho^{\frac{\xi^2}{1-\xi^2}-1} e^{\frac{\xi}{1-\xi^2}u},\\
\eta_{mm} &= \frac{\xi^2}{(1-\xi^2)^2}\rho^{\frac{\xi^2}{1-\xi^2}-1} e^{\frac{\xi}{1-\xi^2}u} > 0,
\end{aligned} \tag{3.5}$$

and

$$\eta_{\rho\rho}\eta_{mm} - \eta_{\rho m}^2 = \frac{\xi^4}{(1-\xi^2)^3}\rho^{\frac{2\xi^2}{1-\xi^2}-2} e^{\frac{2\xi}{1-\xi^2}u} > 0, \tag{3.6}$$

which indicates that η is strictly convex for any $\xi \in (-1,1)$. This implies that

$$\eta(\rho^\varepsilon, m^\varepsilon)_t + q(\rho^\varepsilon, m^\varepsilon)_x$$

is compact in H_{loc}^{-1} due to Diperna [9, 10]. Therefore we have the following lemma.

LEMMA 3.2. *Assume that* $(\rho^\varepsilon, m^\varepsilon)$ *are the solutions of* (3.1), (3.2); *then for any* $\xi \in (-1,1)$,

$$\eta_t(\rho^\varepsilon, m^\varepsilon) + q_x(\rho^\varepsilon, m^\varepsilon) \text{ is compact in } H_{loc}^{-1}, \tag{3.7}$$

where (η, q) *is defined as in* (2.8).

4. Convergence of approximate solutions. This section is devoted to the existence of a weak solution of isothermal gas dynamics. Choose $(\rho^\varepsilon, m^\varepsilon)$ as in (3.4); there exists a subsequence of $(\rho^\varepsilon, m^\varepsilon)$ (still denoted by $(\rho^\varepsilon, m^\varepsilon)$) such that, as $\varepsilon \to 0$,

$$\rho^\varepsilon(x,t) \rightharpoonup \rho(x,t), \quad m^\varepsilon(x,t) \rightharpoonup m(x,t) \tag{4.1}$$

in $L^\infty((0,T)\times R)$ weak star for some measurable functions $\rho(x,t), m(x,t)$.

Let us denote $\nu_{x,t}$ to be the Young measure associated to the weak limits (4.1). For any two entropy pairs in (2.8),

$$\eta_1 = w^{\frac{1}{2(1-\xi_1)}} z^{\frac{1}{2(1+\xi_1)}}, \qquad \eta_2 = w^{\frac{1}{2(1-\xi_2)}} z^{\frac{1}{2(1+\xi_2)}},$$
$$q_1 = (u+\xi_1)\eta_1, \qquad q_2 = (u+\xi_2)\eta_2, \qquad \xi_1, \xi_2 \in (-1,1),$$

Lemma 3.2 gives

$$\begin{aligned}\langle \nu_{x,t}, q_1\eta_2 - q_2\eta_1\rangle = {} & \langle \nu_{x,t}, q_1\rangle\langle \nu_{x,t}, \eta_2\rangle \\ & -\langle \nu_{x,t}, q_2\rangle\langle \nu_{x,t}, \eta_1\rangle \quad \text{for almost every} \quad x,t,\end{aligned} \tag{4.2}$$

i.e.,

$$\begin{aligned}(\xi_1-\xi_2)\langle \nu_{x,t}, \eta_1\eta_2\rangle = {} & \langle \nu_{x,t}, (u+\xi_1)\eta_1\rangle\langle \nu_{x,t}, \eta_2\rangle \\ & -\langle \nu_{x,t}, (u+\xi_2)\eta_2\rangle\langle \nu_{x,t}, \eta_1\rangle \quad \text{for almost every} \quad x,t.\end{aligned} \tag{4.3}$$

We shall show that $\nu_{x,t}$ is either a point mass or concentrated in the vacuum. To this end, we show that (4.2) holds for any $\xi_1, \xi_2 \in \mathbf{C}$, except the two points $(-1,0)$ and $(1,0)$, through analytic extension theorem. In other words, we establish the commutation relations for both weak and strong waves.

Since w, z is bounded, it is convenient to study (4.2) in the $w-z$ plane. Let

$$\Omega = \{(w,z); 0 \le w_- \le w \le w_+,\ 0 \le z_- \le z \le z_+\}$$

be the smallest rectangle containing the support of a fixed $\nu_{x,t}$. It is easy to see, if $w_+ = 0$ or $z_+ = 0$, that Ω is supported in the vacuum. So we assume that $w_- < w_+$, $z_- < z_+$ in what follows.

Let $\xi_1 = 1 - \frac{1}{2n}$; then $\eta_1 = \eta_n = w^n z^{\frac{n}{4n-1}}$. From (4.2), we have

$$\langle \nu_{x,t}, q_n\eta - q\eta_n\rangle = \langle \nu_{x,t}, q_n\rangle\langle \nu_{x,t}, \eta\rangle - \langle \nu_{x,t}, q\rangle\langle \nu_{x,t}, \eta_n\rangle. \tag{4.4}$$

We define probability measure μ_z as follows: for any $h \in C_0(R^2)$,

$$\langle \mu_z, h\rangle = \lim_{n\to\infty} \frac{\langle \nu_{x,t}, hw^n z^{\frac{n}{4n-1}}\rangle}{\langle \nu_{x,t}, w^n z^{\frac{n}{4n-1}}\rangle}. \tag{4.5}$$

Similar to [9], it is easy to check that the support of the probability measure μ_z is contained in the line $w = w_+$.

On the other hand, (4.4) yields

$$\frac{\langle \nu_{x,t}, q_n\eta - q\eta_n\rangle}{\langle \nu_{x,t}, \eta_n\rangle} = \frac{\langle \nu_{x,t}, q_n\rangle}{\langle \nu_{x,t}, \eta_n\rangle}\langle \nu_{x,t}, \eta\rangle - \langle \nu_{x,t}, q\rangle. \tag{4.6}$$

Let $n \to \infty$; then we have

$$\langle \mu_z, q - \lambda_2\eta\rangle = \langle \nu_{x,t}, q - \lambda_2^+\eta\rangle, \tag{4.7}$$

where $\lambda_2^+ = \langle \mu_z, \lambda_2 \rangle$ is finite even though $\lambda_2 = u + 1 = \frac{\ln w - \ln z}{2} + 1$ may go to infinity. In fact, since the left term of (4.7) is finite and $\langle \nu_{x,t}, \eta \rangle$ is positive, λ_2^+ must be bounded. In the same way, let $\xi = -1 + \frac{1}{2n}$; then we have

$$\langle \mu_w, q - \lambda_1 \eta \rangle = \langle \nu_{x,t}, q - \lambda_1^+ \eta \rangle, \tag{4.8}$$

with

$$\begin{aligned} &\langle \mu_w, h \rangle = \lim_{n\to\infty} \frac{\langle \nu_{x,t}, h w^{\frac{n}{4n-1}} z^n \rangle}{\langle \nu_{x,t}, w^{\frac{n}{4n-1}} z^n \rangle} \quad \forall h \in C_0(R^2). \\ &\lambda_1^+ = \langle \mu_w, \lambda_1 \rangle. \end{aligned} \tag{4.9}$$

Combining (4.7) and (4.8), we have the following lemma.

LEMMA 4.1. *Let* $\eta = w^{\frac{1}{2(1-\xi)}} z^{\frac{1}{2(1+\xi)}}, q = (u + \xi)\eta, \xi \in (-1, 1)$*; then the following holds:*

$$\begin{aligned} (\lambda_2^+ - \lambda_1^+) \frac{\langle \nu_{x,t}, \eta \rangle}{\eta(w^+, z^+)} &= (1 + \xi) \left\langle \mu_w, \left(\frac{w}{w^+} \right)^{\frac{1}{2(1-\xi)}} \right\rangle \\ &\quad + (1 - \xi) \left\langle \mu_z, \left(\frac{z}{z^+} \right)^{\frac{1}{2(1+\xi)}} \right\rangle. \end{aligned} \tag{4.10}$$

In view of Lemma 4.1, we can prove $\nu_{x,t}(\{\rho > 0\}) = 1$ for almost every x, t. This means that there is no positive mass concentrated in the vacuum for the measure $\nu_{x,t}$. In fact, by the definitions of μ_w and μ_z, it is easy to check that $\mu_w(\{w > 0\}) = \mu_z(\{z > 0\}) = 1$, and thus (4.10) can be rewritten as

$$\begin{aligned} (\lambda_2^+ - \lambda_1^+) \frac{\langle \nu_{x,t}, \eta \rangle}{\eta(w^+, z^+)} &= \left\langle \mu_w, (1 + \xi) \left[\left(\frac{w}{w^+} \right)^{\frac{1}{2(1-\xi)}} - 1 \right] \right\rangle \\ &\quad + \left\langle \mu_z, (1 - \xi) \left[\left(\frac{z}{z^+} \right)^{\frac{1}{2(1+\xi)}} - 1 \right] \right\rangle + 2. \end{aligned} \tag{4.11}$$

Letting $\operatorname{Im} \xi \to \infty$, we get

$$\begin{aligned} (\lambda_2^+ - \lambda_1^+) \nu_{x,t}(\{\rho > 0\}) &= -\frac{1}{2} \left\langle \mu_w, \ln \left(\frac{w}{w^+} \right) \right\rangle - \frac{1}{2} \left\langle \mu_z, \ln \left(\frac{z}{z^+} \right) \right\rangle + 2 \\ &= (\lambda_2^+ - \lambda_1^+) \end{aligned} \tag{4.12}$$

due to $\langle \mu_w, \ln w_+ \rangle = \langle \mu_z, \ln w_+ \rangle = \ln w_+$ and $0 \le \frac{w}{w_+}, \frac{z}{z_+} \le 1$. This implies $\nu_{x,t}(\{\rho > 0\}) = 1$.

Since the functions $\langle \nu_{x,t}, \eta \rangle$, $\langle \mu_w, w^{\frac{1}{2(1-\xi)}} \rangle$, and $\langle \mu_z, z^{\frac{1}{2(1+\xi)}} \rangle$ are analytic in the domain Ω_w and (4.10) holds in the segment $(-1, 1)$, (4.10) must hold for all $\xi \in \Omega_w = \{\xi;\ -1 < \operatorname{Re} \xi < 1\}$ due to analytic extension theorem. We now establish the commutation relations for the whole complex space except the two singular points $(-1, 0)$ and (1,0); i.e., (4.10) holds not only for weak entropies but also for the strong ones.

For any function $h \in C_0(R^2)$, we define a probability measure as follows:

$$\langle \overline{\mu}_z, h \rangle = \lim_{n\to\infty} \frac{\langle \nu_{x,t}, h w^n \rangle}{\langle \nu_{x,t}, w^n \rangle}. \tag{4.13}$$

In the same way as in [9], it is easy to check supp $\overline{\mu}_z \subset \{(w,z); w = w^+\}$. We compute that, for all $h \in C_0(R^2)$,

$$\begin{aligned}\langle \mu_z, h\rangle &= \lim_{n\to\infty} \frac{\langle \nu_{x,t}, hw^n z^{\frac{n}{4n-1}}\rangle}{\langle \nu_{x,t}, w^n\rangle} \lim_{n\to\infty} \frac{\langle \nu_{x,t}, w^n\rangle}{\langle \nu_{x,t}, w^n z^{\frac{n}{4n-1}}\rangle} \\ &= \frac{\langle \overline{\mu}_z, hz^{\frac{1}{4}}\rangle}{\langle \overline{\mu}_z, z^{\frac{1}{4}}\rangle},\end{aligned} \tag{4.14}$$

which implies that $\langle \mu_z, z^{\frac{1}{2(1+\xi)}}\rangle$ is analytic in the domain Re $\frac{1}{2(1+\xi)} > -\frac{1}{4}$. In the same way, $\langle \mu_w, w^{\frac{1}{2(1-\xi)}}\rangle$ is also analytic in the domain Re $\frac{1}{2(1-\xi)} > -\frac{1}{4}$. Thus, the right term of (4.10) is analytic in the domain $\Omega_0 = \{\xi \in \mathbf{C}; \mathrm{Re}\frac{1}{2(1+\xi)} > -\frac{1}{4}, \quad \mathrm{Re}\frac{1}{2(1-\xi)} > -\frac{1}{4}\}$. Next we show that (4.10) holds for any $\xi \in \bar{\Omega}_0 = \{\xi \in \mathbf{C}; |\xi| > 3\} \subset \Omega_0$ by the analytic extension theorem. Since $\eta = w^{\frac{1}{2(1-\xi)}} z^{\frac{1}{2(1+\xi)}}$ may be unbounded in $\bar{\Omega}_0$, it is necessary to show that $\langle \nu_{x,t}, \eta\rangle$ is well defined for any $\xi \in \bar{\Omega}_0$.

We compute

$$\begin{aligned} f(\xi; w, z) &= \left(\frac{w}{w_+}\right)^{\frac{1}{2(1-\xi)}} \left(\frac{z}{z_+}\right)^{\frac{1}{2(1+\xi)}} + \left(\frac{w}{w_+}\right)^{\frac{1}{2(1+\xi)}} \left(\frac{z}{z_+}\right)^{\frac{1}{2(1-\xi)}} \\ &= \mathrm{e}^{\frac{\ln\frac{w}{w_+}+\ln\frac{z}{z_+}}{2(1-\xi^2)}} \left[\mathrm{e}^{\frac{(\ln\frac{w}{w_+}-\ln\frac{z}{z_+})\xi}{2(1-\xi^2)}} + \mathrm{e}^{\frac{-(\ln\frac{w}{w_+}-\ln\frac{z}{z_+})\xi}{2(1-\xi^2)}}\right] \\ &= 2\left\{1 + \sum_{k=1}^{\infty} \frac{\left(-\ln\frac{w}{w_+} - \ln\frac{z}{z_+}\right)^k}{2^k k!} \left(\frac{\frac{1}{\xi^2}}{1-\frac{1}{\xi^2}}\right)^k\right\} \\ &\quad \times \left\{1 + \sum_{j=1}^{\infty} \frac{\left(\ln\frac{w}{w_+} - \ln\frac{z}{z_+}\right)^{2j}}{4^j (2j)!} \left(\frac{\frac{1}{\xi}}{1-\frac{1}{\xi^2}}\right)^{2j}\right\} \\ &= 2\left\{1 + \sum_{m=1}^{\infty} \left[\sum_{k=1}^{m} \frac{\left(-\ln\frac{w}{w_+} - \ln\frac{z}{z_+}\right)^k}{2^k k!} \binom{m-1}{k-1}\right] \frac{1}{\xi^{2m}}\right\} \\ &\quad \times \left\{1 + \sum_{l=1}^{\infty} \left[\sum_{j=1}^{l} \frac{\left(\ln\frac{w}{w_+} - \ln\frac{z}{z_+}\right)^{2j}}{4^j (2j)!} \binom{l+j-1}{2j-1}\right] \frac{1}{\xi^{2l}}\right\} \\ &= \sum_{p=0}^{\infty} \frac{c_{2p}(w,z)}{\xi^{2p}}. \end{aligned} \tag{4.15}$$

It is observed that $\ln\frac{w}{w_+} + \ln\frac{z}{z_+} \le 0$, we have $c_{2p}(w,z) \ge 0$ in supp $\nu_{x,t}$.

From (4.10), we have

$$
\begin{aligned}
(\lambda_2^+ - \lambda_1^+)\langle \nu_{x,t}, f(\xi; w, z)\rangle &= \langle \mu_w, f(\xi; w, z)\rangle + \langle \mu_z, f(\xi; w, z)\rangle \\
&\quad + \left\langle \mu_w, \xi \left[\left(\frac{w}{w_+}\right)^{\frac{1}{2(1-\xi)}} - \left(\frac{w}{w_+}\right)^{\frac{1}{2(1+\xi)}} \right] \right\rangle \\
&\quad + \left\langle \mu_z, \xi \left[\left(\frac{z}{z_+}\right)^{\frac{1}{2(1-\xi)}} - \left(\frac{z}{z_+}\right)^{\frac{1}{2(1+\xi)}} \right] \right\rangle .
\end{aligned} \tag{4.16}
$$

Similar to (4.15), we expand the last two terms of (4.16) to the Laurent series. We compute

$$
\begin{aligned}
g(\xi; w) &= \xi \left[\left(\frac{w}{w_+}\right)^{\frac{1}{2(1-\xi)}} - \left(\frac{w}{w_+}\right)^{\frac{1}{2(1+\xi)}} \right] \\
&= \xi \left[\mathrm{e}^{\frac{\ln \frac{w}{w_+}}{2(1-\xi^2)}} \left(\mathrm{e}^{\frac{\xi \ln \frac{w}{w_+}}{2(1-\xi^2)}} - \mathrm{e}^{\frac{-\xi \ln \frac{w}{w_+}}{2(1-\xi^2)}} \right) \right] \\
&= \left\{ 1 + \sum_{m=1}^{\infty} \left[\sum_{k=1}^{m} \frac{\left(-\ln \frac{w}{w_+}\right)^k}{2^k k!} \begin{pmatrix} m-1 \\ k-1 \end{pmatrix} \right] \frac{1}{\xi^{2m}} \right\} \\
&\quad \times \left\{ \sum_{l=0}^{\infty} \left[\sum_{j=0}^{l} \frac{-\left(\ln \frac{w}{w_+}\right)^{2j+1}}{4^j (2j+1)!} \begin{pmatrix} l+j \\ 2j \end{pmatrix} \right] \frac{1}{\xi^{2l}} \right\} \\
&= \sum_{p=0}^{\infty} \frac{h_{2p}(\frac{w}{w_+})}{\xi^{2p}}.
\end{aligned} \tag{4.17}
$$

It is obvious that $h_{2p}(\frac{w}{w_+}) \geq 0$. We note that $\langle \mu_w, f(\xi; w, z)\rangle$, $\langle \mu_w, g(\xi; w)\rangle$, $\langle \mu_z, g(\xi; z)\rangle$, and $\langle \mu_z, f(\xi; w, z)\rangle$ are analytic in Ω_0 due to (4.14). Thus $\langle \mu_w, c_{2p}(w, z)\rangle$, $\langle \mu_w, h_{2p}(\frac{w}{w_+})\rangle$, $\langle \mu_z, h_{2p}(\frac{z}{z_+})\rangle$, and $\langle \mu_z, c_{2p}(w, z)\rangle$ are bounded. Now we choose $\xi \in R$ and $|\xi| > 3$; then Levi's lemma gives

$$
\begin{aligned}
\langle \mu_w, f(\xi; w, z)\rangle &= \sum_{p=0}^{\infty} \frac{\langle \mu_w, c_{2p}(w, z)\rangle}{\xi^{2p}}, \\
\langle \mu_z, f(\xi; w, z)\rangle &= \sum_{p=0}^{\infty} \frac{\langle \mu_z, c_{2p}(w, z)\rangle}{\xi^{2p}}, \\
\langle \mu_w, g(\xi; w)\rangle &= \sum_{p=0}^{\infty} \frac{\langle \mu_w, h_{2p}(\frac{w}{w_+})\rangle}{\xi^{2p}}, \\
\langle \mu_z, g(\xi; z)\rangle &= \sum_{p=0}^{\infty} \frac{\langle \mu_z, h_{2p}(\frac{z}{z_+})\rangle}{\xi^{2p}}
\end{aligned} \tag{4.18}
$$

due to the fact that $c_{2p}(w,z), h_{2p}(\frac{w}{w_+}), h_{2p}(\frac{z}{z_+}) \geq 0$. This yields that

$$\sum_{p=0}^{\infty} \frac{\langle \mu_w, c_{2p}(w,z)\rangle}{\xi^{2p}} \tag{4.19}$$

is absolutely convergent for any $\xi \in \bar{\Omega}_0$ because $\langle \mu_w, f(\xi; w, z)\rangle$ is analytic in $\bar{\Omega}_0$. In the same way,

$$\sum_{p=0}^{\infty} \frac{\langle \mu_z, c_{2p}(w,z)\rangle}{\xi^{2p}}, \quad \sum_{p=0}^{\infty} \frac{\langle \mu_w, h_{2p}(\frac{w}{w_+})\rangle}{\xi^{2p}}, \quad \text{and} \quad \sum_{p=0}^{\infty} \frac{\langle \mu_z, h_{2p}(\frac{z}{z_+})\rangle}{\xi^{2p}} \tag{4.20}$$

are also convergent in $\bar{\Omega}_0$.

Thus, from (4.10) and (4.16)–(4.20), we have

$$\begin{aligned}(\lambda_2^+ - \lambda_1^+)\langle \nu_{x,t}, f(\xi; w, z)\rangle = &\sum_{p=0}^{\infty} \frac{\langle \mu_w, c_{2p}(w,z) + h_{2p}(\frac{w}{w_+})\rangle}{\xi^{2p}} \\ &+ \sum_{p=0}^{\infty} \frac{\langle \mu_z, c_{2p}(w,z) + h_{2p}(\frac{z}{z_+})\rangle}{\xi^{2p}}\end{aligned} \tag{4.21}$$

if $\xi \in \Omega_w$. We define

$$\begin{aligned}g_p^{(n)}(w,z) &= (-1)^p n^{2p}\left(f(ni; w, z) - \sum_{s=0}^{p-1}(-1)^s \frac{c_{2s}(w,z)}{n^{2s}}\right), \\ &= \sum_{s=p}^{\infty}(-1)^{p+s}\frac{c_{2s}(w,z)}{n^{2(s-p)}}, \qquad p = 1, 2, \ldots .\end{aligned} \tag{4.22}$$

It is easy to see that $g_p^{(n)}(w,z) \geq 0$, and it converges to $c_{2p}(w,z)$ for almost every $\nu_{x,t}$ as $n \to \infty$. We note that $c_0(w,z) = 2$, and $\langle \nu_{x,t}, c_0(w,z)\rangle$ is well defined. Furthermore, direct computation yields

$$\begin{aligned}&(\lambda_2^+ - \lambda_1^+)\langle \nu_{x,t}, c_0(w,z)\rangle \\ &= \langle \mu_w, c_0(w,z) + h_0(\tfrac{w}{w_+})\rangle + \langle \mu_z, c_0(w,z) + h_0(\tfrac{z}{z_+})\rangle.\end{aligned} \tag{4.23}$$

Now we assume that $\langle \nu_{x,t}, c_{2s}(w,z)\rangle$ are well defined and

$$\begin{aligned}&(\lambda_2^+ - \lambda_1^+)\langle \nu_{x,t}, c_{2s}(w,z)\rangle \\ &= \left\langle \mu_w, c_{2s}(w,z) + h_{2s}\left(\frac{w}{w_+}\right)\right\rangle \\ &\quad + \left\langle \mu_z, c_{2s}(w,z) + h_{2s}\left(\frac{z}{z_+}\right)\right\rangle\end{aligned} \tag{4.24}$$

for $s = 0, 1, \ldots, p-1$. We shall show $\langle \nu_{x,t}, c_{2p}(w,z)\rangle$ is also well defined and (4.24) holds for $s = p$.

From (4.21), (4.22), and (4.24), it is easy to check that $\lim_{n\to\infty}\langle \nu_{x,t}, g_p^{(n)}(w,z)\rangle$ exists. By Fatou's lemma, we have

$$\langle \nu_{x,t}, c_{2p}(w,z)\rangle \leq \lim_{n\to\infty}\langle \nu_{x,t}, g_p^{(n)}(w,z)\rangle, \tag{4.25}$$

which, together with (4.24), implies

$$\begin{aligned}&(\lambda_2^+ - \lambda_1^+)\langle \nu_{x,t}, c_{2p}(w,z)\rangle \\ &\le \left\langle \mu_w, c_{2p}(w,z) + h_{2p}\left(\frac{w}{w_+}\right)\right\rangle + \left\langle \mu_z, c_{2p}(w,z) + h_{2p}\left(\frac{z}{z_+}\right)\right\rangle.\end{aligned} \tag{4.26}$$

To prove (4.24) for $s = p$, we use the way of contradiction. We assume that

$$\begin{aligned}&(\lambda_2^+ - \lambda_1^+)\langle \nu_{x,t}, c_{2p}(w,z)\rangle \\ &< \left\langle \mu_w, c_{2p}(w,z) + h_{2p}\left(\frac{w}{w_+}\right)\right\rangle + \left\langle \mu_z, c_{2p}(w,z) + h_{2p}\left(\frac{z}{z_+}\right)\right\rangle.\end{aligned} \tag{4.27}$$

It is observed that

$$g_p^{(n)}(w,z) = c_{2p}(w,z) - \frac{1}{n^2} g_{p+1}^{(n)}(w,z). \tag{4.28}$$

Thus, we have

$$\begin{aligned}&(\lambda_2^+ - \lambda_1^+)\left\langle \nu_{x,t}, \frac{1}{n^2} g_{p+1}^{(n)}(w,z)\right\rangle \\ &= (\lambda_2^+ - \lambda_1^+)\langle \nu_{x,t}, c_{2p}(w,z)\rangle - \left\langle \mu_w, c_{2p}(w,z) + h_{2p}\left(\frac{w}{w_+}\right)\right\rangle \\ &\quad - \left\langle \mu_z, c_{2p}(w,z) + h_{2p}\left(\frac{z}{z_+}\right)\right\rangle \\ &\quad + \sum_{s=p+1}^{\infty} (-1)^{p+s+1} \frac{\langle \mu_w, c_{2s}(w,z) + h_{2s}(\frac{w}{w_+})\rangle}{\xi^{2(s-p)}} \\ &\quad + \sum_{s=p+1}^{\infty} (-1)^{p+s+1} \frac{\langle \mu_z, c_{2s}(w,z) + h_{2s}(\frac{z}{z_+})\rangle}{\xi^{2(s-p)}},\end{aligned} \tag{4.29}$$

which yields $\lim_{n\to\infty}\langle \nu_{x,t}, \frac{1}{n^2} g_{p+1}^{(n)}(w,z)\rangle < 0$. This contradicts the fact that $g_{p+1}^{(n)}(w,z) > 0$.

By the induction principle, $\langle \nu_{x,t}, c_{2p}(w,z)\rangle$ are well defined and (4.24) holds for all $p = 0, 1, \ldots$. Therefore the Laurent series

$$\sum_{p=0}^{\infty} \frac{\langle \nu_{x,t}, c_{2p}(w,z)\rangle}{\xi^{2p}} \tag{4.30}$$

is convergent for any $\xi \in \bar{\Omega}_0$.

Now we construct a sequence of $\nu_{x,t}$-measurable functions

$$f_m(\xi; w, z) = \sum_{p=0}^{m} \frac{c_{2p}(w,z)}{\xi^{2p}} \tag{4.31}$$

which converges to $f(\xi; w, z)$ in $\{w > 0, z > 0\}$. Since

$$\langle \nu_{x,t}, |f_m(\xi; w, z)|\rangle \le \sum_{p=0}^{\infty} \frac{\langle \nu_{x,t}, c_{2p}(w,z)\rangle}{|\xi|^{2p}} < \infty \tag{4.32}$$

for $\xi \in \bar{\Omega}_0$, again using Fatou's lemma, we have $f(\xi; w, z) \in \mathbf{L}(d\nu_{x,t})$ if $\xi \in \bar{\Omega}_0$. In particular, $w^{-\alpha} z^{\frac{\alpha}{4\alpha+1}} + z^{-\alpha} w^{\frac{\alpha}{4\alpha+1}} \in \mathbf{L}(d\nu_{x,t})$ for any $0 < \alpha < \frac{1}{4}$. Therefore, again using the analytic extension theorem, (4.10) holds in the domain $\bar{\Omega}_0$. We note that η is exactly a strong entropy if $\xi \in \bar{\Omega}_0/\Omega_w$; thus we establish (4.10) for some strong entropies.

Choosing any constant $\alpha \in (0, \frac{1}{4})$, we define a probability measure μ_{1z} like (4.13); i.e., for any $h \in C_0(R^2)$,

$$\langle \mu_{1z}, h\rangle = \lim_{n\to\infty} \frac{\langle \nu_{x,t}, h w^n z^{-\alpha}\rangle}{\langle \nu_{x,t}, w^n z^{-\alpha}\rangle}. \tag{4.33}$$

We compute

$$\begin{aligned} \langle \mu_z, h\rangle &= \lim_{n\to\infty} \frac{\langle \nu_{x,t}, h w^n z^{\frac{n}{4n-1}}\rangle}{\langle \nu_{x,t}, w^n z^{-\alpha}\rangle} \lim_{n\to\infty} \frac{\langle \nu_{x,t}, w^n z^{-\alpha}\rangle}{\langle \nu_{x,t}, w^n z^{\frac{n}{4n-1}}\rangle} \\ &= \frac{\langle \mu_{1z}, h z^{\frac{1}{4}+\alpha}\rangle}{\langle \mu_{1z}, z^{\frac{1}{4}+\alpha}\rangle}. \end{aligned} \tag{4.34}$$

In terms of previous argument, we have (4.10) for any $\xi \in \bar{\Omega}_1 = \{\xi \in \mathbf{C}; |\xi| > \frac{2}{1+4\alpha} + 1\} \subset \Omega_1 = \{\xi \in \mathbf{C}; \mathrm{Re}\frac{1}{2(1+\xi)} > -\frac{1}{4} - \alpha \quad \text{and} \quad \mathrm{Re}\frac{1}{2(1-\xi)} > -\frac{1}{4} - \alpha\}$. Thus, repeating the above arguments, we can establish (4.10) for any $\xi \in \bar{\Omega}_k = \{\xi \in \mathbf{C}; |\xi| > \frac{2}{1+4k\alpha} + 1\} \subset \Omega_k = \{\xi \in \mathbf{C}; \mathrm{Re}\frac{1}{2(1+\xi)} > -\frac{1}{4} - k\alpha \quad \text{and} \quad \mathrm{Re}\frac{1}{2(1-\xi)} > -\frac{1}{4} - k\alpha\}$, where k is an arbitrary positive integer. Therefore we have the following lemma.

LEMMA 4.2. *For any $\xi_1, \xi_2 \in C/\{(-1,0),(0,1)\}$, the following holds:*

$$\begin{aligned} \langle \nu_{x,t}, q_1\eta_2 - q_2\eta_1\rangle = &\langle \nu_{x,t}, q_1\rangle\langle \nu_{x,t}, \eta_2\rangle \\ &-\langle \nu_{x,t}, q_2\rangle\langle \nu_{x,t}, \eta_1\rangle \quad \textit{for almost every} \quad x, t, \end{aligned} \tag{4.35}$$

where $\eta_i, q_i, i = 1, 2$, are chosen as in (2.8).

Remark 4.3. Lemma 4.2 indicates that the Tartar commutation relations hold for both weak and strong waves chosen in (2.8).

Proof of Theorem 1.4. To prove Theorem 1.4, it is sufficient to show that $\nu_{x,t}$ is a point mass.

Let $\xi_1 = 1 - \frac{1}{2n}$ and $\xi_2 = 1 + \frac{1}{2n}$ in (4.35). We then have

$$\frac{\langle \nu_{x,t}, q(\xi_1)\eta(\xi_2) - q(\xi_2)\eta(\xi_1)\rangle}{\langle \nu_{x,t}, \eta(\xi_1)\rangle\langle \nu_{x,t}, \eta(\xi_2)\rangle} = \frac{\langle \nu_{x,t}, q(\xi_1)\rangle}{\langle \nu_{x,t}, \eta(\xi_1)\rangle} - \frac{\langle \nu_{x,t}, q(\xi_2)\rangle}{\langle \nu_{x,t}, \eta(\xi_2)\rangle}. \tag{4.36}$$

Letting $n \to \infty$ gives

$$\lim_{n\to\infty} \frac{\langle \nu_{x,t}, q(\xi_2)\rangle}{\langle \nu_{x,t}, \eta(\xi_2)\rangle} = \lambda_2^+. \tag{4.37}$$

On the other hand, let $\xi \in \Omega_w, \xi_2 = 1 + \frac{1}{2n}$ in (4.35). Similar to (4.6) and (4.7), we have, as $n \to \infty$,

$$\langle \mu_{-z}, q - \lambda_2\eta\rangle = \langle \nu_{x,t}, q - \lambda_2^+\eta\rangle, \tag{4.38}$$

due to (4.37), where the probability measure μ_{-z} is defined for any $h \in C_0(R^2)$, by

$$\langle \mu_{-z}, h\rangle = \lim_{n\to\infty} \frac{\langle \nu_{x,t}, h w^{-n} z^{\frac{n}{4n+1}}\rangle}{\langle \nu_{x,t}, w^{-n} z^{\frac{n}{4n+1}}\rangle}. \tag{4.39}$$

It is easy to see that the support of the measure μ^{-z} is contained in the line $\{w = w_-\}$.

By (4.7) and (4.38), we have

$$\langle \mu_z, q - \lambda_2 \eta \rangle = \langle \mu_{-z}, q - \lambda_2 \eta \rangle \tag{4.40}$$

for any $\xi \in \Omega_w$.

Let $\eta = w^n z^{\frac{n}{4n-1}}, q = (u + 1 - \frac{1}{2n})\eta$. We then compute

$$w_+^n \langle \mu_z, z^{\frac{n}{4n-1}} \rangle = w_-^n \langle \mu_{-z}, z^{\frac{n}{4n-1}} \rangle, \tag{4.41}$$

which implies $w_- = w_+$ as $n \to \infty$. In the same way, we also have $z_- = z_+$. Thus $\nu_{x,t}$ is either a point mass or supported in the vacuum. This indicates Theorem 1.4 due to the standard theory of compensated compactness.

Proof of Theorem 1.2. Choose $(\rho^\varepsilon, m^\varepsilon)$ as in (3.4); then Lemma 3.1 gives, for any test function $\phi \in C_0^\infty(R_+^2)$,

(4.42)

$$\iint_{t>0} (\rho^\varepsilon \phi_t + m^\varepsilon \phi_x)\, dxdt + \int_R \rho_0^\varepsilon(x)\phi(x,0)\, dx = -\iint_{t>0} \varepsilon \rho^\varepsilon \phi_{xx} dxdt,$$

$$\iint_{t>0} m^\varepsilon \phi_t + \left(\frac{(m^\varepsilon)^2}{\rho^\varepsilon} + \rho^\varepsilon \right) \phi_x\, dxdt + \int_R m_0^\varepsilon(x)\phi(x,0)\, dx = -\iint_{t>0} \varepsilon m^\varepsilon \phi_{xx} dxdt.$$

By Theorem 1.4, there exists a strong convergent subsequence of $(\rho^\varepsilon, m^\varepsilon)$ (still denoted by $(\rho^\varepsilon, m^\varepsilon)$) such that

$$(\rho^\varepsilon(x,t), m^\varepsilon(x,t)) \to (\rho(x,t), m(x,t)) \quad \text{a.e.} \tag{4.43}$$

Letting $\varepsilon \to 0$, (1.4) holds from (4.42). Thus $(\rho(x,t), m(x,t))$ is a weak solution of (1.1) and (1.3) with $\gamma = 1$. Since $(\rho(x,t), m(x,t))$ is the limit of the viscosity solutions $(\rho^\varepsilon, m^\varepsilon)$, it is easy to check that (1.5) holds for any weak convex entropy. Therefore we complete the proof of Theorem 1.2.

Acknowledgments. The authors wish to thank Prof. Xiaqi Ding for kindly suggesting this problem. The authors are also very grateful to Prof. Guiqiang Chen for kindly pointing out a gap in our first proofs. Finally, the first author wishes to thank Prof. Alberto Bressan for his kind hospitality.

REFERENCES

[1] F. Bouchut, S. Jin, and X. Li, *Numerical approximations of pressureless and isothermal gas dynamics*, SIAM J. Numer. Anal., to appear.

[2] G. Q. Chen, *The Theory of Compensated Compactness and the System of Isentropic Gas Dynamics*, Preprint MCS-P154-0590, University of Chicago, 1990.

[3] G. Q. Chen, *Remarks on Diperna's paper "Convergence of the viscosity method for isentropic gas dynamics,"* Proc. Amer. Math. Soc., 125 (1997), pp. 2981–2986.

[4] G. Q. Chen and P. Le Floch, *Compressible Euler equations with general pressure law*, Arch. Ration. Mech. Anal., 153 (2000), pp. 221–259.

[5] K. N. Chueh, C. C. Conley, and J. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 372–411.

[6] X. Q. Ding, G. Q. Chen, and P. Z. Luo, *Convergence of the Lax-Friedrichs scheme for isentropic gas dynamics.* I, Acta Math. Sci. (English Ed.), 5 (1985), pp. 415–432.

[7] X. Q. Ding, G. Q. Chen, and P. Z. Luo, *Convergence of the Lax-Friedrichs scheme for isentropic gas dynamics.* II, Acta Math. Sci. (English Ed.), 5 (1985), pp. 433–472.

[8] X. Q. Ding, G. Q. Chen, and P. Z. Luo, *Convergence of the fractional step Lax-Friedrichs scheme and Godunov scheme for isentropic gas dynamics*, Comm. Math. Phys., 121 (1989), pp. 63–84.
[9] R. Diperna, *Convergence of approximate solutions to conservation laws*, Arch. Ration. Mech. Anal., 82 (1983), pp. 27–70.
[10] R. Diperna, *Convergence of viscosity method for isentropic gas dynamics*, Comm. Math. Phys., 91 (1983), pp. 1–30.
[11] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Commun. Pure Appl. Math., 18 (1965), pp. 697–715.
[12] P. D. Lax, *Hyperbolic systems of conservation laws*, II, Commun. Pure Appl. Math., 10 (1957), pp. 537–566.
[13] P. D. Lax, *Shock waves and entropy*, in Contributions to Nonlinear Functional Analysis, E. A. Zarantonello, ed., Academic Press, New York, 1971, pp. 603–644.
[14] P. L. Lions, B. Perthame, and E. Tadmor, *Kinetic formulation of the isentropic gas dynamics and p-systems*, Comm. Math. Phys., 163 (1994), pp. 169–172.
[15] P. L. Lions, B. Perthame, and P. Souganidis, *Existence and stability of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates*, Commun. Pure Appl. Math., 49 (1996), pp. 599–638.
[16] Y. G. Lu, *An Explicit Lower Bound of Viscosity Solutions to Isentropic Gas Dynamics and to Euler Equations*, Preprint 95-18, SFB 359, Heidelberg University, Germany, 1995.
[17] F. Murat, *Compacité par compensation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 5 (1978), pp. 489–507.
[18] T. Nishida, *Global solution for an initial-boundary value problem of a quasilinear hyperbolic systems*, Proc. Japan. Acad., 44 (1968), pp. 642–646.
[19] T. Nishida and J. Smoller, *Solutions in the large for some nonlinear hyperbolic conservation laws*, Commun. Pure Appl. Math., 26 (1973), pp. 183–200.
[20] F. Poupaud, M. Rascle, and J. Vila, *Global solutions to the isothermal Euler-Poisson system with arbitrarily large data*, J. Differential Equations, 123 (1995), pp. 93–121.
[21] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1982.
[22] L. Tartar, *Compensated compactness and applications to partial equations*, in Nonlinear Analysis and Mechanics, Heriot-Watt Symposium, R. J. Knops, ed., Res. Notes in Math. 4, Pitman, London, 1979, pp. 136–212.

SIAM J. MATH. ANAL.
Vol. 34, No. 3, pp. 611–635

ON A FREE BOUNDARY PROBLEM FOR A STRONGLY DEGENERATE QUASI-LINEAR PARABOLIC EQUATION WITH AN APPLICATION TO A MODEL OF PRESSURE FILTRATION*

RAIMUND BÜRGER†, HERMANO FRID‡, AND KENNETH H. KARLSEN§

Abstract. We consider a free boundary problem of a quasi-linear strongly degenerate parabolic equation arising from a model of pressure filtration of flocculated suspensions. We provide definitions of generalized solutions of the free boundary problem in the framework of L^2 divergence-measure fields. The formulation of boundary conditions is based on a Gauss–Green theorem for divergence-measure fields on bounded domains with Lipschitz deformable boundaries and avoids referring to traces of the solution. This allows one to consider generalized solutions from a larger class than BV. Thus it is not necessary to derive the usual uniform estimates of spatial *and* time derivatives of the solutions of the corresponding regularized problem, as required by the BV approach. We first prove the existence and uniqueness of the solution of the regularized parabolic free boundary problem and then apply the vanishing viscosity method to prove the existence of a generalized solution to the degenerate free boundary problem.

Key words. free boundary problem, strongly degenerate parabolic equation, divergence-measure field, pressure filtration

AMS subject classifications. 35K65, 35R35

PII. S0036141002401007

1. Introduction.

Conventional analyses of initial-boundary value problems of strongly degenerate parabolic equations, which includes first-order conservation laws, are usually based on the concept of generalized solutions in $BV(Q_T)$, where $Q_T := \Omega \times [0,T]$, $\Omega \subset \mathbb{R}$, is the computational domain (for simplicity, assumed to be cylindrical here) [2, 4, 5, 25, 26, 27]. To prove that a generalized solution u of a conservation law or of a strongly degenerate parabolic equation belongs to $BV(Q_T)$, it is necessary to derive estimates of $\|\partial_x u_\varepsilon\|_{L^1(Q_T)}$ and $\|\partial_t u_\varepsilon\|_{L^1(Q_T)}$ which are uniform with respect to the regularization parameter ε, where u_ε denotes the smooth solution of the corresponding regularized initial-boundary value problem. These estimates (and a uniform L^∞ bound on u_ε) imply that the family $\{u_\varepsilon\}_{\varepsilon>0}$ is compact in $L^1(Q_T)$; i.e., there exists a sequence $\varepsilon = \varepsilon_n$ with $\varepsilon_n \to 0$ for $n \to \infty$ such that $\{u^{\varepsilon_n}\}$ converges in $L^1(Q_T)$ to a limit $u \in L^\infty(Q_T) \cap BV(Q_T)$. It is usually straightforward to verify that this limit is indeed a generalized solution.

The importance of the choice of the space $BV(Q_T)$ lies in the existence of traces of the limit function u with respect to the lateral boundaries of Q_T. This well-known property of BV functions is stated, e.g., in [11, sect. 5.32, Thm. 1]. As has become

*Received by the editors January 16, 2002; accepted for publication (in revised form) May 31, 2002; published electronically January 7, 2003. This work was supported by the Sonderforschungsbereich 404 at the University of Stuttgart and by the Applied Mathematics in Industrial Flow Problems (AMIF) programme of the European Science Foundation (ESF).
http://www.siam.org/journals/sima/34-3/40100.html

†Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany (buerger@mathematik.uni-stuttgart.de).

‡Instituto de Matemática Pura e Aplicada (IMPA), Estrada Dona Castorina 110, Jardim Botânico, CEP 22460-320, Rio de Janeiro, RJ, Brazil (hermano@impa.br). The work of this author was supported by the CNPq through grants 352871/96-2 and 46.5714/00-5, and by the FAPERJ through grant E-26/151.890/2000.

§Department of Mathematics, University of Bergen, Johs. Brunsgt. 12, N-5008 Bergen, Norway (kenneth.karlsen@mi.uib.no).

apparent in [4], traces are needed in the proof of uniqueness of generalized solutions.

For several reasons, the BV approach unfortunately imposes some severe limitations on the analysis of initial-boundary value problems of hyperbolic and strongly degenerate parabolic equations. The most obvious one is the apparent difficulty in actually deriving the required uniform estimates of $\|\partial_x u_\varepsilon\|_{L^1(Q_T)}$ and $\|\partial_t u_\varepsilon\|_{L^1(Q_T)}$. This worked out, e.g., for the spatially one-dimensional problems analyzed in [4]. However, for only marginally more involved equations (but still in one space dimension), and in particular for different boundary conditions, it seems no longer possible to derive a uniform estimate of $\|\partial_t u_\varepsilon\|_{L^1(Q_T)}$. An example of such an initial-boundary problem is given in [24]. When passing to several space dimensions, i.e., to equations of the type

$$\partial_t u + \nabla_{\mathbf{x}} \cdot \mathbf{f}(u) = \Delta A(u), \quad (\mathbf{x}, t) \in Q_T := \Omega \times [0, T], \quad \Omega \subset \mathbb{R}^n, \tag{1.1}$$

together with initial and boundary conditions and where the function $A(u)$ is nonnegative, increasing, and Lipschitz continuous, it seems virtually impossible to derive the required uniform estimates, where the estimate on the spatial derivative has to be replaced, of course, by a uniform estimate of $\|\nabla_{\mathbf{x}} u_\varepsilon\|_{L^1(Q_T)}$.

In the cases where only a uniform estimate of $\|\nabla_{\mathbf{x}} u_\varepsilon\|_{L^1(Q_T)}$ (but not of the time derivative) is feasible, one can utilize Kružkov's "interpolation lemma" [14, Lem. 5] in order to conclude that the sequence u_ε converges to a limit function u belonging to the wider class $BV_{1,1/2}(Q_T) \supset BV(Q_T)$. This means that there exists a constant K such that

$$\iint_{Q_T} \big|u(\mathbf{x} + \Delta\mathbf{x}, t) - u(\mathbf{x}, t)\big| d\mathbf{x}dt \le K|\Delta\mathbf{x}|,$$
$$\iint_{Q_T} \big|u(\mathbf{x}, t + \Delta t) - u(\mathbf{x}, t)\big| d\mathbf{x}dt \le K|\Delta t|^{1/2}.$$

Note that the $BV_{1,1/2}$ estimates of $\{u_\varepsilon\}$ are entirely sufficient to apply Kolmogorov's compactness criterion in order to show existence of a limit function. The problem is with boundary conditions and uniqueness, since it is not ensured that a function $u \in BV_{1,1/2}(Q_T)$ possesses traces at the boundaries of Q_T, such that boundary conditions need to be defined in a fashion that avoids these traces; however, it is then not obvious how to prove uniqueness.

Another general limitation of the BV approach has become apparent in [4] and is due to the restriction that the initial datum u_0 of that paper belongs to the class

$$\mathcal{B} := \Big\{ u \in BV(\Omega) : \ u(x) \in \mathcal{U}_0 \ \forall x \in \overline{\Omega};\ \mathrm{TV}_\Omega(\partial_x A_\varepsilon(u)) < M_0 \text{ uniformly in } \varepsilon \Big\},$$

where $A'_\varepsilon(u) = a_\varepsilon(u)$ and a_ε is an appropriately regularized, positive diffusion coefficient. The condition $u_0 \in \mathcal{B}$ is required to ensure that $\|\partial_t u_\varepsilon(\cdot, t)\|_{L^1(\Omega)}$ or $\|\partial_t u_\varepsilon\|_{L^1(Q_T)}$ remain uniformly bounded. For a given, in general discontinuous function u_0, membership in $\mathcal{B}$ is difficult to verify due to the discontinuity of the diffusion coefficient $a(u)$, so $\mathcal{B}$ denotes a possibly very narrow class.

The mentioned difficulties associated with the BV approach make it desirable to consider generalized solutions from a wider class. This wider class is associated here with the notion of divergence-measure fields, which is a class of vector fields that was first considered by Anzellotti [1]. This paper is based on the recent formulation by Chen and Frid [9].

The main idea is to replace the requirement $u \in L^\infty(Q) \cap BV(Q)$, where we consider $Q \subset \mathbb{R}^N$ and which can be expressed as

$$\|u\|_{BV(Q)} < \infty, \quad \|u\|_{BV(Q)} := \sup\left\{ \int_Q u \nabla \cdot \varphi \, d\mathbf{x} : \varphi \in \left(C_0^1(Q)\right)^N, \ \|\varphi\|_{L^\infty(Q)} \le 1 \right\},$$

by the requirement that a vector field $F \in L^p(Q, \mathbb{R}^N)$ associated with the sought solution u satisfy

$$|\operatorname{div} F|(Q) < \infty, \quad |\operatorname{div} F|(Q) := \sup\left\{ \int_Q F \cdot \nabla \varphi \, d\mathbf{x} : \varphi \in C_0^1(Q; \mathbb{R}), \ \|\varphi\|_{L^\infty(Q)} < 1 \right\}.$$

We define the class of L^p divergence-measure vector fields over Q by

$$\mathcal{DM}^p(Q) := \left\{ F \in L^p(Q; \mathbb{R}^N) : \ |\operatorname{div} F|(Q) < \infty \right\}.$$

We see that if $F \in \mathcal{DM}^p(Q)$, then $\operatorname{div} F$ is a Radon measure over Q. If we assume that the components of F are Lipschitz continuous functions of u, as in the application to conservation laws (see below), then it becomes clear that $u \in L^\infty(Q) \cap BV(Q)$ implies $F \in \mathcal{DM}^\infty(Q)$.

Properties of divergence-measure fields for the case $p = \infty$ are derived by Chen and Frid in [9]. Most important, it is possible to prove a generalized Gauss–Green formula for divergence-measure fields in bounded domains using the concept of domains with deformable Lipschitz boundaries, which allows the definition of traces. For the case of scalar conservation laws, the importance of divergence-measure fields accrues from the fact that any convex entropy pair actually forms an L^∞ divergence-measure field over $Q \subset \mathbb{R}^N$ if we consider a bounded spatial domain $\Omega \subset \mathbb{R}^{N-1}$. Utilizing the Gauss–Green formula, Chen and Frid [9] provide an appropriate formulation for L^∞ (not BV) solutions of conservation laws with boundary conditions. They are able to derive a formulation of an entropy boundary condition which was proposed previously by Otto [17, 19, 20, 21] by advancing the concept of entropy boundary fluxes.

Most properties of L^p, $p = \infty$, divergence-measure fields derived in [9] also hold for $1 \le p < \infty$, as is detailed in [10]. The case $p = 2$ is of particular interest for the analysis of degenerate parabolic equations, since in light of standard a priori estimates, it is possible to show that the appropriately defined entropy pair of a strongly degenerate parabolic equation is an L^2 divergence-measure field over $Q_T \subset \mathbb{R}^{N-1} \times [0, T]$. (More general domains can be considered, but we may limit the discussion here to cylindrical domains.) This was first exploited in a recent paper by Mascia, Porretta, and Terracina [18], who proved existence and uniqueness of L^∞ solutions to nonhomogeneous Dirichlet initial-boundary value problems of (1.1), which in particular includes entropy boundary conditions.

In [6] entropy boundary conditions for strongly degenerate parabolic equations in the context of an application to sedimentation with compression are derived. However, the definition of traces of the solution with respect to the lateral boundary of the computational domain is only possible if the diffusion coefficient $a(u)$ is, for example, Lipschitz continuous. This assumption does not hold for the cases we are interested in here. Moreover, although Dirichlet boundary conditions in the context of solid-liquid separation models lead to mathematically well-posed initial-boundary value problems, their physical significance is questionable due to violation of a conservation principle. Rather, kinematic "flux-type" or "wall" boundary conditions (such as that

of Problem B of [4]) should be employed. In fact, it turned out that these boundary conditions are satisfied in an a.e. pointwise sense on the lateral boundaries of Q_T, that is, in a much stronger sense than are entropy boundary conditions, although they also involve the concept of traces.

The above discussion motivates our interest in applying the recently developed divergence-measure theory to initial-boundary value problems of strongly degenerate parabolic equations. We could now treat again the initial-boundary value problems studied, e.g., in [4] in an appropriate divergence-measure framework and obtain an existence and uniqueness result. However, since the BV calculus is indeed applicable to those problems, the chief gain in using the more general divergence-measure concept would merely consist of the relaxation of the condition $u_0 \in \mathcal{B}$. Instead, the theory of L^2 divergence-measure fields is applied here to a free boundary problem, which is a slight modification of a model of pressure filtration presented in [3]. The problem is still one-dimensional, and its boundary conditions are of "flux-type," similar to those of [4]. Since the flux contains the derivative of the degenerate parabolic term which is only bounded in L^2, we cannot consider strong traces for this term. Moreover, there is reason to believe that the mentioned BV estimate of $\partial_t u_\varepsilon$ cannot be derived. This conjecture is based on the observation that in many other analyses it was necessary to differentiate the corresponding regularized viscous equation with respect to t, to multiply it with a suitable sign-type function, and to use integration by parts. The problem with the filtration problem is the occurrence of the derivative (with respect to t) of the free boundary as a coefficient in the equation, such that differentiating the entire equation with respect to t would entail the necessity to estimate $h''(t)$. Due to the coupling condition with the solution evaluated at one of the boundaries, however, we have no control over this quantity. This seems to preclude the necessary uniform estimate of $\partial_t u$.

The remainder of this paper is organized as follows. In section 2 we briefly recall the mathematical model of pressure filtration, state the free boundary problem, and provide a brief definition of L^2 divergence-measure fields together with the properties relevant for the subsequent analysis. In section 3 generalized solutions of the free boundary problem are defined, where an equivalent problem transformed to fixed boundaries is also considered. In section 4 we state the corresponding regularized viscous free boundary problems and show that they have a unique solution for fixed values of the regularization parameter. Finally we conclude in section 5 by the viscosity method that there exists a generalized solution to the free boundary problem in the sense of section 3.

The analysis of the free boundary problem has not yet been completed, since a uniqueness proof is still lacking. It is, however, not obvious, for instance, how the uniqueness proof for a comparable free boundary problem by Zhao and Li [28], which is based on establishing a fixed boundary initial-boundary value problem for a suitably complemented generalized solution of the free boundary problem, can be extended to the free boundary problem studied in this paper.

2. Statement of the problem and preliminaries.

2.1. Pressure filtration of flocculated suspensions. To motivate the free boundary problem, we briefly recall the one-dimensional mathematical model of pressure filtration formulated in [3]. We consider a filter column closed at height $z = 0$ by a filter medium, which lets only the liquid pass, and at a variable height $z = h(t)$ by a piston which moves downwards due to an applied pressure $\sigma(t)$. The material behavior of the suspension is described by two model functions, the flux density function

or hindered settling factor f and the effective solid stress function σ_e, both functions only of the local solids concentration u. Here f is a nonpositive Lipschitz continuous function with compact support in $[0, u_{\max}]$, where $u_{\max} \le 1$ is the maximum concentration, and the function σ_e satisfies $\sigma_e = 0$ for $u \le u_c$, where $0 \le u_c \le u_{\max}$ is a critical concentration value, and $\sigma_e'(u) > 0$ for $u > u_c$. According to the phenomenological sedimentation-consolidation theory [3, 7, 8], the evolution of the concentration distribution is given by the equation

$$\partial_t u + \partial_z\big(h'(t)u + f(u)\big) = \partial_z^2 A(u), \quad 0 \le z \le h(t), \quad 0 < t \le T, \tag{2.1}$$
$$A(u) := \int_0^u a(s)\,ds, \quad a(u) := Cu^{-1} f(u)\sigma_e'(u),$$

where the parameter $C < 0$ expresses the solid-fluid density difference. Observe that (2.1) is hyperbolic for $u \le u_c$ and $u \ge u_{\max}$ and parabolic for $u_c < u < u_{\max}$ and thus of strongly degenerate parabolic type since the degeneration to hyperbolic type takes place on an interval of solution values of positive length. Specifically for the filtration problem, we assume that the solids flux through the moving piston and through the filter medium is zero. Since (2.1) is derived from the solids continuity equation, this implies the kinematic boundary conditions

$$\big(f(u) - \partial_z A(u)\big)\big(h(t), t\big) = 0, \quad \big(h'(t)u + f(u) - \partial_z A(u)\big)(0, t) = 0, \quad t > 0.$$

At time $t = 0$, the column is filled with a suspension of the local initial volumetric concentration $u(z, 0) = u_0(z)$ for $0 \le z \le h(0) := 1$.

The salient mathematical difficulty of the pressure filtration model arises from the coupling between the applied pressure $\sigma = \sigma(t)$ and the piston trajectory $h(t)$. Resistance to the movement of the piston, i.e., to the flow rate of filtrate leaving the filter, is exerted by the filter medium and by the so-called filter cake forming above the medium. While the resistance of the filter medium is constant, that of the filter cake depends on its thickness and composition, that is, on the solution u. The growth of the filter cake during the initial stages of the filtration process therefore slows down the downward movement of the piston if the applied pressure is kept constant. Specifically, a vertical stress balance and an application of Darcy's law yield the following coupling equation between $\sigma(t)$ and $h(t)$ [3, 16], which is written here as an ordinary differential equation for h:

$$h'(t) + \beta(t)h(t) + \gamma\big(t, u(0,t)\big) = 0, \quad 0 < t \le T, \tag{2.2}$$
$$\beta(t) := \frac{g\varrho_f}{\mu_f R}, \quad \gamma\big(t, u(0,t)\big) := \frac{1}{\mu_f R}\big[g\big(m_0 - \varrho_f\big) + \sigma(t) - \sigma_e\big(u(0,t)\big)\big].$$

Here g is the acceleration of gravity, ϱ_f the density of the fluid, μ_f its viscosity, R the resistance of the filter medium, and m_0 the initial suspension mass divided by the cross-sectional area of the filter column.

The observation that γ depends on $\sigma_e(u(0,t))$ and not on some arbitrary function of $u(0,t)$ is essential to making the problem amenable to mathematical analysis. In fact, both functions σ_e and A vanish for $u \le u_c$, strictly increase for $u_c < u < u_{\max}$, and remain constant for $u \ge u_{\max}$. Thus we can express $\sigma_e(u)$ as a function of $A(u)$, and the function γ takes the form

$$\gamma\big(t, u(0,t)\big) = \tilde{\gamma}(t) + \alpha\big(A(u(0,t))\big), \tag{2.3}$$

where α is a monotonous function on $[u_c, u_{\max}]$ having an inverse α^{-1}.

For numerical examples and applications to experimental data we refer to [3, 12].

2.2. Statement of the free boundary problem. A natural property of any solution u of the free boundary problem in the context of the pressure filtration model should be $0 \le u \le 1$; i.e., solution values should be physically relevant as concentration values. However, due to the presence of the linear transport term $h'(t)u$ in combination with the kinematic boundary condition prescribed at $z = 0$, we cannot exclude that boundary layers involving nonphysical solution values form. This can be avoided if we consider that from a physical point of view, the piston stops immediately as soon as the filter is "clogged," i.e., when the solid particles at $z = 0$ form a dense packing. We consider this effect by replacing the coupling condition (2.2) by the condition

$$h'(t) + c\big(A(u(0,t))\big)\big[\beta(t)h(t) + \gamma\big(t, u(0,t)\big)\big] = 0, \quad 0 < t \le T, \tag{2.4}$$

where $c(\rho) = 1$ for $\rho \in (0, A(u_{\max}))$ and $c(\rho) = 0$ otherwise.

Finally, we introduce a new space coordinate $x = h(t) - z$. Then $x = 0$ corresponds to the piston and $x = h(t)$ to the filter medium. Observing that $\partial_t(u(x,t)) = \partial_t u(z,t) + h'(t)\partial_z u$ and replacing $f(u)$ by $-f(u)$, we get the following free boundary value problem:

$$\partial_t u + \partial_x f(u) = \partial_x^2 A(u), \quad (x,t) \in Q(h,T), \tag{2.5a}$$

$$u(x,0) = u_0(x), \quad 0 \le x \le 1, \tag{2.5b}$$

$$\big(f(u) - \partial_x A(u)\big)(0,t) = 0, \quad 0 < t \le T, \tag{2.5c}$$

$$\big(f(u) - \partial_x A(u)\big)\big(h(t),t\big) = h'(t)u\big(h(t),t\big), \quad 0 < t \le T, \tag{2.5d}$$

$$h'(t) + c\Big(A\big(u\big(h(t),t\big)\big)\Big)\Big[\beta(t)h(t) + \gamma\Big(t, u\big(h(t),t\big)\Big)\Big] = 0, \quad 0 < t \le T, \tag{2.5e}$$

$$h(0) = 1, \tag{2.5f}$$

where $Q(h,T) := \{(x,t) \in (0,1) \times (0,T] : 0 < x < h(t)\}$.

Also, after the change of variables above, the relation (2.3) becomes

$$\gamma\big(t, u(h(t),t)\big) = \tilde{\gamma}(t) + \alpha\big(A(u(h(t),t))\big). \tag{2.6}$$

Since we are interested here exclusively in solutions that take values in the interval $[0, u_{\max}] \subset [0,1]$ of admissible concentrations, we may assume that $a(u) = 0$ for $u \le u_c$ and $u \ge u_{\max}$ such that $A(u) = A(u_{\max})$ for $u \ge u_{\max}$ and $A(u) = 0$ for $u \le u_c$. In particular, we have $0 = \alpha(0) \le \alpha(A(u(0,t))) \le \alpha(A(u_{\max})) =: K_\alpha$ for all times. Since, moreover, $\widetilde{\gamma}$ is a control function given a priori, we may assume that there exist positive constants $k_{\tilde{\gamma}}$ and $K_{\tilde{\gamma}}$ with $k_{\tilde{\gamma}} \le \widetilde{\gamma}(t) \le K_{\tilde{\gamma}}$ for all $t \in [0,T]$ and thus that there exist $k_\gamma, K_\gamma > 0$ with $k_\gamma \le \gamma \le K_\gamma$ for all $t \in [0,T]$. Similarly, we may assume that there exist $k_\beta, K_\beta > 0$ with $k_\beta \le \beta(t) \le K_\beta$ for all $t \in [0,T]$. Finally, to establish well-posedness of the free boundary problem, we assume that $T < 1/K_\gamma$.

2.3. Divergence-measure fields. Here we briefly recall the basic facts of the theory of divergence-measure fields as developed in [9, 10]. Since we will be interested only in the L^2 divergence-measure fields, we will focus our discussion on that case.

Let $\Omega \subset \mathbb{R}^N$ be an open bounded subset. We denote by $\mathcal{DM}^2(\Omega)$ the space of all $L^2(\Omega)$ vector fields whose divergence is a bounded Radon measure on Ω:

$$\mathcal{DM}^2(\Omega) := \left\{ F \in (L^2(\Omega))^N : \exists C > 0 : \forall \varphi \in C_0^\infty(\Omega), \ \left| \int_\Omega F \cdot \nabla\varphi \, d\mathbf{x} \right| \le C\|\varphi\|_\infty \right\},$$

where, as usual, $C_0^\infty(\Omega)$ denotes the space of the infinitely differentiable functions with compact support contained in Ω. Analogously, one may define $\mathcal{DM}^p(\Omega)$, $1 \le p \le \infty$, replacing L^2 by L^p, and $\mathcal{DM}^{\rm ext}(\Omega)$ replacing $L^2(\Omega)^N$ by $\mathcal{M}(\Omega)^N$, the space of vector-valued Radon measures over Ω with N components.

DEFINITION 2.1. *We say that $\partial\Omega$ is a* deformable Lipschitz boundary *provided that the following hold:*

(a) *For all $x \in \partial\Omega$ there exists a number $r > 0$ and a Lipschitz map $h : \mathbb{R}^{N-1} \to \mathbb{R}$ such that, after rotating and relabeling coordinates if necessary,*

$$\Omega \cap Q(x,r) = \left\{ y \in \mathbb{R}^N : h(y_1, \dots, y_{N-1}) < y_N \right\} \cap Q(x,r),$$

where $Q(x,r) := \{ y \in \mathbb{R}^N : |x_i - y_i| \le r,\ i = 1, \dots, N \}$.

(b) *There exists a mapping $\Psi : \partial\Omega \times [0,1] \to \overline{\Omega}$ such that Ψ is a homeomorphism bi-Lipschitz over its image and $\Psi(\omega, 0) = \omega$ for all $\omega \in \partial\Omega$. The map Ψ is called a* Lipschitz deformation *of the boundary $\partial\Omega$. We denote $\Psi_s(\omega) = \Psi(\omega, s)$ and $\partial\Omega_s = \Psi_s(\partial\Omega)$. We also denote by Ω_s the bounded open set whose boundary is $\partial\Omega_s$.*

The following theorem is a particular case of a general result proved in [10], following the guidelines in [9]; we refer to [10] for the proof. If $\mathcal{C}$ is a closed set, we denote by $\mathrm{Lip}(\mathcal{C})$ the space of Lipschitz functions defined on $\mathcal{C}$, equipped with the norm $\|f\|_{\rm Lip} = \|f\|_\infty + \mathrm{Lip}(f)$.

THEOREM 2.2. *Let $F \in \mathcal{DM}^2(\Omega)$, with Ω a bounded open set with Lipschitz deformable boundary. Then there exists a continuous linear functional $F \cdot \nu|_{\partial\Omega}$ over $\mathrm{Lip}(\partial\Omega)$ such that, for any $\phi \in \mathrm{Lip}(\mathbb{R}^N)$,*

$$\langle F \cdot \nu|_{\partial\Omega}, \phi|_{\partial\Omega} \rangle = \int_\Omega \phi \,\mathrm{div}\, F + \int_\Omega \nabla\phi \cdot F. \tag{2.7}$$

Moreover, let $\nu : \Psi(\partial\Omega \times [0,1]) \to \mathbb{R}^N$ be such that $\nu(x)$ is the outer unit normal to $\partial\Omega_s$ at $x \in \partial\Omega_s$, defined for almost every $x \in \Psi(\partial\Omega \times [0,1])$. Then, for any $\psi \in \mathrm{Lip}(\partial\Omega)$,

$$\langle F \cdot \nu|_{\partial\Omega}, \psi \rangle = \operatorname*{ess\,lim}_{s \to 0} \frac{1}{s} \int_0^s \left(\int_{\partial\Omega_s} \mathcal{E}(\psi) F \cdot \nu \, d\mathcal{H}^{N-1} \right) ds,$$

where $\mathcal{E}(\psi)$ denotes any Lipschitz extension of ψ to all $\mathbb{R}^N$ and $\mathcal{H}^{N-1}$ is the $(N-1)$-dimensional Hausdorff measure.

As an example, below we will consider a domain Ω of the form

$$\Omega = \{ (x,t) \in \mathbb{R}^2 : 0 < x < h(t),\ 0 < t < T \},$$

where h is a nonincreasing Lipschitz function satisfying $h(t) > h_0$ for some constant $h_0 > 0$. Clearly, in this case Ω satisfies (a) of Definition 2.1. We may also easily define a Lipschitz deformation for $\partial\Omega$. Indeed, since Ω is convex, given any point (x_*, t_*) in its interior, we may define the map $\Psi((x,t), s) = (x + s\delta(x_* - x), t + s\delta(t_* - t))$, from $\partial\Omega \times [0,1]$ to $\overline{\Omega}$, which, for $\delta > 0$ sufficiently small, certainly gives a Lipschitz deformation. But we will prefer to use deformations which, given $\delta > 0$ sufficiently small, on $\{(x,t) : x = 0,\ \delta < t < T - \delta\}$ are given by $\Psi_\delta((0,t), s) = (\delta s, t)$, and on $\{(x,t) : x = h(t),\ \delta < t < T - \delta\}$ are given by $\Psi_\delta((h(t),t), s) = (h(t) - \delta s, t)$. Clearly, Ψ_δ may be extended to all $\partial\Omega \times [0,1]$ in order to provide Lipschitz deformations

for $\partial\Omega$. By the above theorem, if $F \in \mathcal{DM}^2(\Omega)$ and $\phi \in \mathrm{Lip}(\mathbb{R}^2)$ is such that $\operatorname{supp}\phi \cap \partial\Omega \subset \{x = 0\}$, then, for $\delta > 0$ sufficiently small,

$$\langle F \cdot \nu|_{\partial\Omega}, \phi\rangle = \operatorname*{ess\,lim}_{s\to 0} \frac{1}{s}\int_0^s \left(\int_0^T \phi(\delta s, t) F_1(\delta s, t)\, dt\right) ds. \tag{2.8}$$

On the other hand, if $\phi \in \mathrm{Lip}(\mathbb{R}^2)$ is such that $\operatorname{supp}\phi \cap \partial\Omega \subset \{(h(t), t), 0 < t < T\}$, then, for $\delta > 0$ sufficiently small,

$$\langle F \cdot \nu|_{\partial\Omega}, \phi\rangle = \operatorname*{ess\,lim}_{s\to 0} \frac{1}{s}\int_0^s \left(\int_0^T \phi(h(t) - \delta s, t)(F_1 - h'(t)F_2)(h(t) - \delta s, t)\, dt\right) ds. \tag{2.9}$$

3. Definition of generalized solutions. In what follows let K be a sufficiently large constant, e.g., $K = 2u_{\max}$. As above, for fields $F(x,t) = (F_1(x,t), F_2(x,t))$ defined over domains of $\mathbb{R}^2$, which are distributions on these domains, the operator div is defined as $\operatorname{div} F = \partial_x F_1 + \partial_t F_2$ in the sense of distributions.

DEFINITION 3.1. *A pair of functions (u, h) with $h \in C[0,T]$ and $u \in L^\infty(Q(h,T))$ is called a* generalized solution *of the free boundary problem (2.5) if the following conditions are satisfied:*

(a) *The function $h(\cdot)$ is nonincreasing and Lipschitz continuous on $(0,T)$ with $h(0) = 1$, and there exists a positive constant h_0 such that $h(t) > h_0$.*

(b) *The following regularity properties hold:*

$$A(u) \in L^2\big(0,T; H^1(0, h(\cdot))\big), \tag{3.1}$$

$$\forall k \in \mathbb{R}: \ \Big(\operatorname{sgn}(u-k)\big(f(u) - f(k)\big) - \partial_x\big|A(u) - A(k)\big|, |u-k|\Big) \in \mathcal{DM}^2\big(Q(h,T)\big). \tag{3.2}$$

(c) *The boundary conditions are satisfied in the following sense: For $(F_1, F_2) = (f(u) - \partial_x A(u), u)$, $\delta > 0$ sufficiently small, and every test function $\varphi \in C_0^1(\Pi_T)$, with $\Pi_T = \mathbb{R} \times (0,T)$, we have*

$$\operatorname*{ess\,lim}_{s\to 0} \frac{1}{s}\int_0^s \left(\int_0^T \varphi(\delta s, t) F_1(\delta s, t)\, dt\right) ds = 0, \tag{3.3}$$

$$\operatorname*{ess\,lim}_{s\to 0} \frac{1}{s}\int_0^s \left(\int_0^T \varphi(h(t) - \delta s, t)(F_1 - h'(t)F_2)(h(t) - \delta s, t)\, dt\right) ds = 0. \tag{3.4}$$

(d) *Let $\gamma_{x\to h(t)}A(u)$ denote the trace of $A(u)$ for $x \to h(t)$ in the sense of traces in $L^2(0,T; H^1(0,h(\cdot)))$. Then (2.5e) is satisfied a.e. in $(0,T)$, where in $c(A(u(h(t),t)))$ and in $\gamma(t, u(h(t),t))$, given by (2.6), $A(u(h(t),t))$ must be replaced by $\gamma_{x\to h(t)}A(u)$.*

(e) *The initial condition is valid in the sense that*

$$\lim_{t\to 0}\int_0^{h(t)} \big|u(x,t) - u_0(x)\big|\, dx = 0.$$

(f) *The following entropy inequality is satisfied for all nonnegative test functions $\varphi \in C_0^\infty(Q(h,T))$ and all $k \in \mathbb{R}$:*

$$\iint_{Q(h,T)} \Big\{|u-k|\partial_t\varphi + \operatorname{sgn}(u-k)\big[f(u) - f(k) - \partial_x A(u)\big]\partial_x\varphi\Big\}\, dtdx \geq 0. \tag{3.5}$$

It is convenient to transform the free boundary value problem (2.5) to an equivalent initial-boundary value problem with fixed boundaries by introducing a new space coordinate $\xi := x/h(t)$. Wherever notationally convenient, the argument t in $h(t)$ is omitted, and we denote by h^{-1} the function $1/h(t)$, etc. Then we can rewrite (2.5) as the following initial-boundary value problem with fixed boundaries for $v(\xi,t) := u(h(t)\xi,t)$, where $Q_T := (0,1)\times(0,T)$:

$$\partial_\xi v + h^{-1}h'\big(-\partial_\xi(\xi v)+v\big) + h^{-1}\partial_\xi f(v) = h^{-2}\partial_\xi^2 A(v), \quad (\xi,t)\in Q_T, \tag{3.6a}$$

$$v(\xi,0) = u_0(\xi), \quad \xi\in[0,1], \tag{3.6b}$$

$$\big(f(v) - h^{-1}\partial_\xi A(v)\big)(0,t) = 0, \quad t\in(0,T], \tag{3.6c}$$

$$\big(f(v) - h^{-1}\partial_\xi A(v)\big)(1,t) = h'(t)v(1,t), \quad t\in(0,T], \tag{3.6d}$$

$$h'(t) + c\Big(A\big(v\big(1,t\big)\big)\Big)\Big[\beta(t)h(t) + \gamma\Big(t, v\big(1,t\big)\Big)\Big] = 0, \quad 0<t\le T, \tag{3.6e}$$

$$h(0) = 1, \tag{3.6f}$$

while the relation (2.6) becomes

$$\gamma\big(t, v(1,t)\big) = \tilde\gamma(t) + \alpha\big(A(v(1,t))\big). \tag{3.7}$$

In what follows we use $h' := h'(t)$, $h^{-1} := 1/h(t)$, $h^{-2} := 1/(h(t))^2$, and similar notation for the function $h_\varepsilon(t)$ to be defined below. Moreover, we set $g(v,\xi,t) := -h^{-1}h'(t)\xi v + h^{-1}f(v)$.

The appropriate definition of entropy solution in terms of v is as follows.

DEFINITION 3.2. *A pair of functions (v,h) with $h\in C[0,T]$ and $v\in L^\infty(Q_T)$ is called a* generalized solution *of the transformed free boundary problem* (3.6) *if the following conditions are satisfied:*

(a) *The function $h(\cdot)$ is nonincreasing and Lipschitz continuous on $(0,T)$ with $h(0)=1$, and there exists a positive constant h_0 such that $h(t)>h_0$.*

(b) *The following regularity properties hold:*

$$h^{-2}A(v)\in L^2\big(0,T;H^1(0,1)\big), \tag{3.8}$$

$$\forall k\in\mathbb{R}:\ \Big(\operatorname{sgn}(v-k)\big(g(v,\xi,t)-g(k,\xi,t)\big) - h^{-2}\partial_\xi\big|A(v)-A(k)\big|, |v-k|\Big)\in\mathcal{DM}^2(Q_T). \tag{3.9}$$

(c) *The boundary conditions are satisfied in the following sense: For $(F_1,F_2) = (g(v,\xi,t) - h^{-2}\partial_\xi A(v), v)$, $\delta>0$ sufficiently small, and every test function $\varphi\in C_0^1(\Pi_T)$, we have*

$$\operatorname*{ess\,lim}_{s\to0}\frac{1}{s}\int_0^s\left(\int_0^T\varphi(\delta s,t)F_1(\delta s,t)\,dt\right)ds = 0, \tag{3.10}$$

$$\operatorname*{ess\,lim}_{s\to0}\frac{1}{s}\int_0^s\left(\int_0^T\varphi(1-\delta s,t)F_1(1-\delta s,t)\,dt\right)ds = 0. \tag{3.11}$$

(d) *Let $\gamma_{\xi\to1}A(v)$ denote the trace of $A(v)$ for $\xi\to1$ in the sense of traces in $L^2(0,T;H^1(0,1))$. Then* (3.6e) *is satisfied a.e. in $(0,T)$, where in $c(A(v(1,t)))$ and in $\gamma(t,v(1,t))$, given by* (3.7)*, we must replace $A(v(1,t))$ by $\gamma_{x\to1}A(v)$.*

(e) *The initial condition is valid in the sense that*

$$\lim_{t\to0}\int_0^1\big|v(\xi,t)-u_0(\xi)\big|\,d\xi = 0. \tag{3.12}$$

(f) *The following inequality holds for all nonnegative test functions* $\varphi \in C_0^\infty(Q_T)$ *and all* $k \in \mathbb{R}$*:*

$$\begin{aligned}\iint_{Q_T} \Big\{ |v-k|\partial_t\varphi + \big[\operatorname{sgn}(u-k)\big(g(v,\xi,t)-g(k,\xi,t)\big) \\ - \partial_\xi\big|A(v)-A(k)\big|\big]\partial_\xi\varphi\Big\}\,d\xi dt \ge 0.\end{aligned} \tag{3.13}$$

4. Regularized free boundary problem. As in [4] we prove the existence of entropy solutions by the vanishing viscosity method. To this end, we consider the regularized strictly parabolic free boundary problem

$$\partial_t u_\varepsilon + \partial_x f_\varepsilon(u_\varepsilon) = \partial_x^2 A_\varepsilon(u_\varepsilon), \quad (x,t) \in Q(h_\varepsilon, T), \tag{4.1a}$$

$$u_\varepsilon(x,0) = u_0^\varepsilon(x), \quad 0 \le x \le 1, \tag{4.1b}$$

$$\big(f_\varepsilon(u_\varepsilon) - \partial_x A_\varepsilon(u_\varepsilon)\big)(0,t) = 0, \quad 0 < t \le T, \tag{4.1c}$$

$$\big(f_\varepsilon(u_\varepsilon) - \partial_x A_\varepsilon(u_\varepsilon)\big)\big(h_\varepsilon(t),t\big) = h_\varepsilon'(t) u_\varepsilon\big(h_\varepsilon(t),t\big), \quad 0 < t \le T, \tag{4.1d}$$

(4.1e)

$$h_\varepsilon'(t) + c_\varepsilon\Big(A_\varepsilon\big(u_\varepsilon\big(h_\varepsilon(t),t\big)\big)\Big)\Big[\beta_\varepsilon(t)h_\varepsilon(t) + \gamma_\varepsilon\Big(t, u_\varepsilon\big(h_\varepsilon(t),t\big)\Big)\Big] = 0, \quad 0 < t \le T,$$

$$h_\varepsilon(0) = 1. \tag{4.1f}$$

The regularized functions and initial and boundary data are assumed to satisfy first-order compatibility conditions. Problem (4.1) is equivalent to the following initial-boundary value problem with fixed boundaries for $v_\varepsilon(\xi,t) := u_\varepsilon\big(h_\varepsilon(t)\xi, t\big)$ with $(\xi,t) \in Q_T := (0,1)\times(0,T)$:

$$\partial_t v_\varepsilon + h_\varepsilon^{-1}h_\varepsilon'(t)\big[-\partial_\xi(\xi v_\varepsilon) + v_\varepsilon\big] + h_\varepsilon^{-1}\partial_\xi f_\varepsilon(v_\varepsilon) = h_\varepsilon^{-2}\partial_\xi^2 A_\varepsilon(v_\varepsilon), \quad (\xi,t) \in Q_T, \tag{4.2a}$$

$$v_\varepsilon(\xi,0) = u_0^\varepsilon(\xi), \quad 0 \le \xi \le 1, \tag{4.2b}$$

$$\big(f_\varepsilon(v_\varepsilon) - h_\varepsilon^{-1}\partial_\xi A_\varepsilon(v_\varepsilon)\big)(0,t) = 0, \quad 0 < t \le T, \tag{4.2c}$$

$$\big(f_\varepsilon(v_\varepsilon) - h_\varepsilon^{-1}\partial_\xi A_\varepsilon(v_\varepsilon)\big)(1,t) = h_\varepsilon'(t) v_\varepsilon(1,t), \quad 0 < t \le T, \tag{4.2d}$$

$$h_\varepsilon'(t) + c_\varepsilon\big(A\big(v_\varepsilon(1,t)\big)\big)\Big[\beta_\varepsilon h_\varepsilon(t) + \gamma_\varepsilon\big(t, v_\varepsilon(1,t)\big)\Big] = 0, \quad 0 < t \le T, \tag{4.2e}$$

$$h_\varepsilon(0) = 1. \tag{4.2f}$$

We choose the regularization c_ε such that c_ε is smooth, nonnegative, $c_\varepsilon(\rho) = 1$ for $\varepsilon \le \rho \le A(u_{\max}) - \varepsilon$, and $c_\varepsilon(\rho) = 0$ for $\rho \notin (0, A(u_{\max}))$. We assume that the regularization $f_\varepsilon \ge 0$ is also compactly supported, that $a_\varepsilon(u) \ge \varepsilon$, and that $a_\varepsilon(u) - \varepsilon$ is also compactly supported. We assume $\operatorname{supp} f_\varepsilon \cup \operatorname{supp} c_\varepsilon \subset \overline{\mathcal{U}} = [0, u_{\max}]$ and $\operatorname{supp}(a_\varepsilon - \varepsilon) \subset \overline{\mathcal{U}}$. Moreover, we define $g_\varepsilon(u,\xi,t) := -h_\varepsilon^{-1}h_\varepsilon'\xi u + h_\varepsilon^{-1}f_\varepsilon(u)$ and assume that there exist constants ν_ε, L_ε, and $\tilde{L}$ such that

$$\frac{A_\varepsilon(u) - A_\varepsilon(v)}{u - v} \ge \nu_\varepsilon > 0, \quad \big|g_\varepsilon(u,\xi,t) - g_\varepsilon(v,\xi,t)\big| \le L_\varepsilon|u-v| \quad \text{for } u,v \in \mathbb{R}. \tag{4.3}$$

LEMMA 4.1. *Any solution* u_ε *of the regularized free boundary problem* (4.1) *satisfies* $u_\varepsilon(x,t) \in \overline{\mathcal{U}}$ *for all* $(x,t) \in \overline{Q(h_\varepsilon, T)}$*. Equivalently, any solution* v_ε *of* (4.2) *satisfies* $v_\varepsilon(x,t) \in \overline{\mathcal{U}}$ *for all* $(x,t) \in \overline{Q_T}$*. In particular, there exists a constant* M_0 *independent of* ε *such that for all sufficiently small* $\varepsilon > 0$*,*

$$\|u_\varepsilon\|_{L^\infty(Q(h_\varepsilon,T))} \le M_0. \tag{4.4}$$

Proof. Consider the regularized problem (4.1), perturbed by adding to the right-hand member the term $\lambda N(u_\varepsilon)$, where $\lambda > 0$ and $N(u) = u_{\max}/2 - u$. We may assume h_ε to be a given smooth function, so the problem is in fact given by the first four equations of (4.1), with the first one perturbed. If we prove the result for the perturbed problem, then by the well-known stability for quasi-linear strictly parabolic scalar equations, with respect to coefficients, the desired result will follow sending $\lambda \to 0$. Now, if the result is not true for the perturbed problem, there is a time t_0 at which the solution v_ε leaves $\overline{\mathcal{U}}$ for the first time, that is, $t_0 = \inf\{t : v_\varepsilon(x,t) \notin \overline{\mathcal{U}} \text{ for some } x \in [0,h(t)]\}$. In this case, there exists $x_0 \in [0,h(t_0)]$ such that $u_\varepsilon(x_0,t_0) \in \{0,u_{\max}\}$, say, $u_\varepsilon(x_0,t_0) = u_{\max}$. If $x_0 \in (0,h(t_0))$, as usual, we get a contradiction using that $\partial_x u_\varepsilon = 0$, $\partial_t u_\varepsilon \geq 0$, $\partial_x^2 u_\varepsilon \leq 0$, $a_\varepsilon(u) > 0$, and $N(u_{\max}) < 0$. On the other hand, if $x_0 \in \{0,h(t_0)\}$, using (4.1c)–(4.1e), we again conclude that $\partial_x u_\varepsilon = 0$. Hence, we must again have $\partial_t u_\varepsilon \geq 0$, $\partial_x^2 u_\varepsilon \leq 0$, and so we get a contradiction in the same way. ∎

LEMMA 4.2. *Suppose that $T < 1/K_\gamma$ and that the coefficients of the regularized problem* (4.1) *satisfy compatibility conditions. Then this problem has a unique solution $(u_\varepsilon, h_\varepsilon)$ such that $u_\varepsilon \in C^{2+\alpha,1+\alpha/2}(\overline{Q}(h_\varepsilon,T))$ and $h_\varepsilon \in C^{1+\alpha/2}[0,T]$. Precisely, the function h_ε satisfies the following estimates uniformly in ε:*

$$0 < h_0 \leq h_\varepsilon(t) \leq 1, \quad \|h_\varepsilon'\|_{L^\infty(0,T)} \leq M_h := K_\beta + K_\gamma. \tag{4.5}$$

Proof. Suppose that $(u_\varepsilon, h_\varepsilon)$ with $u_\varepsilon \in C^{2,1}(\overline{Q}(h_\varepsilon,T))$ and $h_\varepsilon \in C^1(0,T)$ is a solution of problem (4.1) or, equivalently, that v_ε satisfies the initial-boundary value problem with fixed boundaries (4.2). In addition, consider for a fixed function $h_\varepsilon \in C^1[0,T]$ the initial-boundary value problem (4.2′) consisting of (4.2a) and the initial and boundary conditions (4.2b)–(4.2d).

The proof of the following lemma is standard and can be found, e.g., in [15, Chap. V].

LEMMA 4.3. *Under the assumptions of Lemma* 4.2*, the solution w_ε of the initial-boundary value problem* (4.2′) *satisfies the following estimates, where the constant K_1 is independent of ε:*

$$0 \leq w_\varepsilon \leq K_1, \quad \|w_\varepsilon\|_{C^\beta(Q_T)} \leq K_2, \quad \|\partial_\xi w_\varepsilon\|_{C^{1,1/2}(\overline{Q_T})} \leq K_2, \quad \|w_\varepsilon\|_{W^{2,1}_\infty(\overline{Q}_T)} \leq K_2.$$

To prove the existence of a solution to problem (4.2), we follow Zhao and Li [28] and use the Schauder fixed point theorem. To this end, define the set

$$H := \big\{h \in C^1(0,T) : \|h'\|_\infty \leq M_h,\ h(0) = 1,\ h \text{ is nonincreasing}\big\},$$

where the constant M_h is defined in (4.5). Note that H is a compact convex set in the Banach space $C^0[0,T]$. Moreover, let $\widehat{\beta}_\varepsilon(t,u) := \chi_\varepsilon(A_\varepsilon(u))\beta_\varepsilon(t)$ and $\widehat{\gamma}_\varepsilon(t,u) := \chi_\varepsilon(u)\gamma_\varepsilon(t,u)$.

LEMMA 4.4. *Let the operator $\mathcal{T} : H \to C^0[0,T]$ be defined by*

$$(\mathcal{T}h)(t) := \exp\Big(\widehat{B}_\varepsilon\big(t; w_\varepsilon(1,\cdot)\big)\Big)\left[1 - \int_0^t \exp\Big(-\widehat{B}_\varepsilon\big(\tau; w_\varepsilon(1,\cdot)\big)\Big)\widehat{\gamma}_\varepsilon\big(\tau, w_\varepsilon(1,\tau)\big)\,d\tau\right],$$

$$\widehat{B}_\varepsilon(t;w) := -\int_0^t \widehat{\beta}_\varepsilon\big(\tau, w(\tau)\big)\,d\tau,$$

where w_ε is the solution of the initial-boundary value problem (4.2′) *corresponding to h. Then $\mathcal{T}h \in H$, i.e., the operator $\mathcal{T}$ maps H into itself.*

Proof. Since we consider a fixed value of the regularization parameter ε, we simplify notation in the remainder of the proof of Lemma 4.2 (including the proofs of Lemmas 4.4 and 4.5) by omitting ε wherever possible.

Obviously, we have $(\mathcal{T}h)(0) = 1$. Since the functions $\widehat{B}(\cdot; w)$ and $\widehat{\gamma}\big(\cdot, w(1,\cdot)\big)$ are smooth, as stated in Lemma 4.3, we see that $\mathcal{T}h \in C^1[0,T]$. Furthermore, we have

$$
\begin{aligned}
(\mathcal{T}h)'(t) = &-\widehat{\beta}\big(t, w(1,t)\big) \exp\Big(\widehat{B}\big(t, w(1,t)\big)\Big) \\
&\times \left[1 - \int_0^t \exp\Big(-\widehat{B}\big(\tau; w(1,\cdot)\big)\Big)\widehat{\gamma}\big(\tau, w(1,\tau)\big)\, d\tau\right] - \widehat{\gamma}\big(t, w(1,t)\big).
\end{aligned} \tag{4.6}
$$

Since $\widehat{\gamma}(t, w(1,t)) \le K_\gamma$ for $\varepsilon > 0$ sufficiently small, the expression in the square brackets in (4.6) is nonnegative, and thus $\mathcal{T}h$ is nonincreasing, if the condition $T < 1/K_\gamma$ is satisfied. Moreover, this assumption implies that $|(\mathcal{T}h)'(t)| \le K_\beta + K_\gamma$. We conclude that indeed $\mathcal{T}h \in H$. ∎

To apply the Schauder fixed point theorem, and thus to show existence of the solution, we have to prove the following lemma.

LEMMA 4.5. *Suppose that* $\{h_n\}_{n\in\mathbb{N}} \subset H$ *and* $\|h_m - h_n\|_{C^0[0,T]} \to 0$ *as* $m, n \to \infty$. *Then* $\|\mathcal{T}h_m - \mathcal{T}h_n\|_{C^0[0,T]} \to 0$ *as* $m, n \to \infty$.

Proof. Assume that $h_n \to h$ uniformly in $[0,T]$. Since $\|h_n'\|_\infty \le M_h$, we can conclude that $h' \in L^\infty[0,T]$ and $h_n' \to h'$ weakly in $L^1[0,T]$. Let w_n and w denote the solutions of the initial-boundary value problem (4.2′) associated with the functions h_n and h, respectively. From Lemma 4.3 it follows that there exist subsequences $\{w_{n_j}\}_{j\in\mathbb{N}}$ and $\{\partial_x w_{n_j}\}_{j\in\mathbb{N}}$ of $\{w_n\}_{n\in\mathbb{N}}$ and $\{\partial_x w_n\}_{n\in\mathbb{N}}$, respectively, converging uniformly on $\overline{Q_T}$ to limit functions $\overline{w}$ and $\overline{w}_x$. Multiplying (4.2a), with v replaced by w_{n_j}, by a test function $\varphi \in C_0^2(Q_T)$, integrating over Q_T, and using integration by parts, we obtain

$$
\iint_{Q_T} \Big\{ w_{n_j}\partial_t\varphi + h_{n_j}^{-1}h_{n_j}' w_{n_j}(\varphi + \xi\partial_\xi\varphi) + \big(h_{n_j}^{-1} f(w_{n_j}) - h_{n_j}^{-2}\partial_\xi A(w_{n_j})\big)\partial_\xi\varphi \Big\}\, d\xi dt = 0.
$$

Letting $j \to \infty$, we get

$$
\iint_{Q_T} \Big\{ \overline{w}\partial_t\varphi + h^{-1}h'\overline{w}(\varphi + \xi\partial_\xi\varphi) + \big(h^{-1} f(\overline{w}) - h^{-2}\partial_\xi A(\overline{w})\big]\partial_\xi\varphi \Big\}\, d\xi dt = 0.
$$

Since solutions of the initial-boundary value problem (4.2′) are unique, we obtain $\overline{w} = w$; hence the sequences $\{w_n\}_{n\in\mathbb{N}}$ and $\{\partial_x w_n\}_{n\in\mathbb{N}}$ converge uniformly on $\overline{Q_T}$. Lemma 4.5 is then an immediate consequence of

$$
\begin{aligned}
&(\mathcal{T}h_n - \mathcal{T}h_m)(t) \\
&= \exp\Big(\widehat{B}\big(t, w(1,\cdot)\big)\Big)\int_0^t \exp\Big(-\widehat{B}\big(\tau, w(1,\cdot)\big)\Big)\Big[\widehat{\gamma}\big(\tau, w_m(1,\tau)\big) - \widehat{\gamma}\big(\tau, w_n(1,\tau)\big)\Big]\, d\tau.
\end{aligned}
$$

∎

We continue with the proof of Lemma 4.2. By Lemma 4.5, $\mathcal{T}$ is a continuous operator on H. We are now in a position to conclude from the Schauder fixed point theorem that $\mathcal{T}$ has a fixed point $h \in H$; in particular $h \in C^{1+\alpha/2}[0,T]$. This also proves the estimates (4.5).

Substituting the fixed point h into the initial-boundary value problem (4.2′) produces a solution $w \in C^{2+\alpha,1+\alpha/2}(Q_T)$ with the property that the pair (w,h) also satisfies the fixed point equation $\mathcal{T}h = h$, which is equivalent to (2.5f). Consequently,

$(v \equiv w, h)$ is a solution of the initial-boundary value problem (4.2), and setting $u(x,t) = v(x/h(t),t)$ produces a solution (u,h) of the regularized free boundary problem (4.1) with $u \in C^{2+\alpha,1+\alpha/2}(\overline{Q}(h,T))$. Thus the existence part of Lemma 4.2 is proved.

We now turn to the uniqueness part. From boundary condition (4.1d) we get

$$\frac{1}{2}h^2(t) = \int_0^t h(s)h'(s)\,ds + \frac{1}{2} = \int_0^t \frac{h(s)}{u}\big(f(u) - \partial_x A(u)\big)\big(h(s),s\big)\,ds + \frac{1}{2}.$$

We now choose a test function $\omega \in C^2(\mathbb{R})$ satisfying $\omega(x) = 0$ for $x \le h_0/2$ and $\omega(x) = 1$ for $x \ge 3h_0/4$. We then get

$$\begin{aligned}
&\int_0^t \frac{h(s)}{u}\big(f(u) - \partial_x A(u)\big)\big(h(s),s\big)\,ds = \iint_{Q(h,t)} \partial_x\left(\frac{x\omega(x)}{u}\big(f(u) - \partial_x A(u)\big)\right) dxds\\
&= \iint_{Q(h,t)} \left\{\big(\omega(x) + x\omega'(x)\big)\frac{f(u) - \partial_x A(u)}{u} + x\omega(x)\partial_x\left(\frac{f(u) - \partial_x A(u)}{u}\right)\right\} dxds\\
&= \iint_{Q(h,t)} \big(\omega(x) + x\omega'(x)\big)\frac{f(u) - \partial_x A(u)}{u}\,dxds\\
&\quad + \iint_{Q(h,t)} x\omega(x)\big(f(u) - \partial_x A(u)\big)\partial_x\Big(\frac{1}{u}\Big)\,dxds + \iint_{Q(h,t)} \frac{x\omega(x)}{u}(-\partial_s u)\,dxds\\
&=: I_1 + I_2 + I_3.
\end{aligned}$$

Defining

$$\widetilde{A}(u) := \int_0^u \frac{a(r)}{r}dr, \quad p(u) := \int_0^u \frac{f'(r)}{r}dr, \quad q(u) := \int_{u_0}^u \frac{f(r)}{r^2}dr,$$

with $u_0 > 0$, we obtain by using integration by parts and the boundary condition

$$\begin{aligned}
I_2 &= \iint_{Q(h,t)} x\omega(x)\partial_x\left(\frac{f(u) - \partial_x A_\varepsilon(u)}{u}\right) dxds\\
&\quad - \iint_{Q(h,t)} \frac{x\omega(x)}{u}\partial_x\big(f(u) - \partial_x A_\varepsilon(u)\big)\,dxds\\
&= \int_0^t h(s)\left(\frac{f(u) - \partial_x A_\varepsilon(u)}{u}\right) ds\\
&\quad - \iint_{Q(h,t)} \big(\omega(x) + x\omega'(x)\big)\big(-p(u) + q(u) - \partial_x \widetilde{A}_\varepsilon(u)\big)\,dxds\\
&\quad + \iint_{Q(h,t)} \frac{x\omega(x)}{u}\partial_s u\,dxds\\
&= \int_0^t h(s)\Big\{-p\big(u(h(s),s)\big) + q\big(u(h(s),s)\big) - \partial_x \widetilde{A}\big(u(h(s),s)\big)\Big\}\,ds\\
&\quad - \iint_{Q(h,t)} \Big\{\big(2\omega'(x) + x\omega''(x)\big)\widetilde{A}(u) + \big(\omega(x) + x\omega'(x)\big)\big(p(u) - q(u)\big)\Big\}\,dxds\\
&\quad + \iint_{Q(h,t)} \frac{x\omega(x)}{u}\partial_s u\,dxds + \int_0^t \widetilde{A}_\varepsilon\big(u\big(h(s),s\big)\big)\,ds.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\frac{1}{2}h^2(t) = \frac{1}{2} + \iint_{Q(h,t)} (\omega + x\omega')\frac{f(u) - \partial_x A(u)}{u}\,dxds \\
&+ \int_0^t h(s)\Big\{-p\big(u(h(s),s)\big) + q\big(u(h(s),s)\big) - \partial_x \widetilde{A}\big(u(h(s),s)\big)\Big\}\,ds \\
&+ \int_0^t \widetilde{A}\big(u(h(s),s)\big)\,ds - \iint_{Q(h,t)} \Big\{(\omega + x\omega')\big(p(u) - q(u)\big) + (2\omega' + x\omega'')\widetilde{A}\Big\}\,dxds.
\end{aligned}$$

Now let (u^1, h^1) and (u^2, h^2) be two solutions of the regularized free boundary problem (4.1). Let

$$t_1 := \max\big\{t \in [0,T] :\ h^1(\tau) = h^2(\tau) \text{ for } \tau \in [0,t]\big\}.$$

We now show that $t_1 = T$. To this end, we first suppose that $t_1 < T$. Without loss of generality, we suppose that $t_1 = 0$. Moreover, define $h^-(t) := \min\{h^1(t), h^2(t)\}$, $h^+(t) := \max\{h^1(t), h^2(t)\}$, $j(t) := 1$ if $h^1(t) > h^2(t)$ and $j(t) := 2$ if $h^1(t) \le h^2(t)$, and $i(t) := 3 - j(t)$. Then we obtain

$$\begin{aligned}
&\frac{1}{2}\big((h^1)^2(t) - (h^2)^2(t)\big) \\
&= \iint_{Q(h^-,t)} (\omega + x\omega')\left[\frac{f(u^1) - \partial_x A(u^1)}{u^1} - \frac{f(u^2) - \partial_x A(u^2)}{u^2}\right] dxds \\
&\quad - \int_0^t (-1)^{j(s)} \int_{h^-(s)}^{h^+(s)} (\omega + x\omega')\frac{f(u^{j(s)}) - \partial_x A(u^{j(s)})}{u^{j(s)}}\,dxds \\
&\quad + \int_0^t \Big\{h^1(s)\Big[-p\big(u^1(h^1(s),s)\big) + q\big(u^1(h^1(s),s)\big) - \partial_x \widetilde{A}\big(u^1(h^1(s),s)\big)\Big] \\
&\qquad\qquad - h^2(s)\Big[-p\big(u^2(h^2(s),s)\big) + q\big(u^2(h^2(s),s)\big) - \partial_x \widetilde{A}\big(u^2(h^2(s),s)\big)\Big]\Big\}\,ds \\
&\quad + \int_0^t \Big\{\widetilde{A}\big(u^1(h^1(s),s)\big) - \widetilde{A}\big(u^2(h^2(s),s)\big)\Big\}\,ds \\
&\quad + \iint_{Q(h^-,t)} \Big\{(\omega + x\omega')\big(-p(u^1) + q(u^1) + p(u^2) - q(u^2)\big) \\
&\qquad\qquad - (2\omega' + x\omega'')\big(\widetilde{A}(u^1) - \widetilde{A}(u^2)\big)\Big\}\,dxds \\
&\quad - \int_0^t (-1)^{j(s)} \int_{h^-(s)}^{h^+(s)} \Big\{\big(\omega(x) + x\omega'(x)\big)\Big(-p\big(u^{j(s)}\big) + q\big(u^{j(s)}\big)\Big) \\
&\qquad\qquad - (2\omega' + x\omega'')\widetilde{A}\big(u^{j(s)}\big)\Big\}\,dxds \\
&=: I_4 + \cdots + I_9.
\end{aligned}$$

We now set $\delta(t) := |h^1(t) - h^2(t)|$. First note that

$$\big|(h^1)^2(t) - (h^2)^2(t)\big| = \big|h^1(t) + h^2(t)\big|\delta(t) \ge M_1\delta(t), \quad M_1 := 2h(0).$$

We now estimate the integrals I_4 to I_9. In light of

$$\begin{aligned} I_4 = &\iint_{Q(h^-.t)} (\omega + x\omega')\left(\frac{f(u^1)}{u^1} - \frac{f(u^2)}{u^2}\right) dxds \\ &- \int_0^t \left\{\widetilde{A}\big(u^1(h^-(s),s)\big) - \widetilde{A}\big(u^2(h^-(s),s)\big)\right\} ds \\ &+ \iint_{Q(h^-,t)} (2\omega' + \omega'')\big(\widetilde{A}(u^1) - \widetilde{A}(u^2)\big)\, dxds \end{aligned}$$

and the inequality

$$\big|\widetilde{A}\big(u^1(h^-(s),s)\big) - \widetilde{A}\big(u^2(h^-(s),s)\big)\big| \le \varepsilon^{-1}\|a\|_\infty \big|u^1(h^-(s),s) - u^2(h^-(s),s)\big|,$$

it is easy to see that there exist constants C_2 and C_3 such that

$$|I_4| \le C_2 \int_0^t \big|u^1(h^-(s),s) - u^2(h^-(s),s)\big| ds + C_3 \int_0^t \int_0^{h^-(t)} \big|u^1(x,s) - u^2(x,s)\big|\, dxds.$$

Next, noting that in light of boundary condition (4.1c)

$$\begin{aligned} &\big|f\big(u^{j(s)}(x,s)\big) - \partial_x A\big(u^{j(s)}(x,s)\big)\big| \\ &= \big|f\big(u^{j(s)}(x,s)\big) - f\big(u^{j(s)}(h^+(s),s)\big) \\ &\qquad - \partial_x A\big(u^{j(s)}(x,s)\big) + \partial_x A\big(u^{j(s)}(h^+(s),s)\big)\big| \\ &\le \Big(\|f'\|_\infty \|\partial_x u(\cdot,s)\|_\infty + \|a'\|_\infty \|\partial_x u(\cdot,s)\|_\infty \\ &\qquad + \|a\|_\infty \|\partial_x^2 u(\cdot,s)\|_\infty\Big)\big|x - h^+(s)\big|, \end{aligned} \tag{4.7}$$

we obtain that there exists a constant C_4 satisfying $|I_5| \le C_4\delta^2(t)$. Observe that

$$\begin{aligned} &\big|\widetilde{A}\big(u^1\big(h^1(s),s\big)\big) - \widetilde{A}\big(u^2\big(h^2(s),s\big)\big)\big| \\ &\le \big|\widetilde{A}\big(u^{j(s)}(h^+(s),s)\big) - \widetilde{A}\big(u^{j(s)}(h^-(s),s)\big)\big| \\ &\qquad + \big|\widetilde{A}\big(u^{j(s)}(h^-(s),s)\big) - \widetilde{A}\big(u^{i(s)}(h^-(s),s)\big)\big| \\ &\le \varepsilon^{-1}\|a\|_\infty \|\partial_x u(\cdot,s)\|_\infty \delta(t) \\ &\qquad + \varepsilon^{-1}\|a\|_\infty \big|u^{j(s)}\big(h^-(s),s\big) - u^{i(s)}\big(h^-(s),s\big)\big|. \end{aligned}$$

From this inequality and similar ones for the functions $\partial_x\widetilde{A}$, p, and q, we obtain that there exist constants C_5 and C_6 such that

$$|I_6| + |I_7| \le C_5 \int_0^t \delta(\tau)\, d\tau + C_6 \int_0^t \big|u^1\big(h^-(s),s\big) - u^2\big(h^-(s),s\big)\big|\, ds.$$

By similar arguments it follows that there exist constants C_7 and C_8 satisfying

$$|I_8| \le C_7 \int_0^t \delta(\tau)\, d\tau + C_8 \int_0^t \int_0^{h^-(s)} \big|u^1(x,s) - u^2(x,s)\big|\, dxds.$$

Finally, since the integrand of I_9 is bounded, there exists a constant C_9 such that

$$|I_9| \le C_9 \int_0^t \delta(\tau)\, d\tau.$$

Summarizing the estimates of I_4 to I_9, we obtain

$$\begin{aligned}\delta(t) \le C_4\delta^2(t) + C_{10}\int_0^t \big|u^1\big(h^-(s),s\big) - u^2\big(h^-(s),s\big)\big|\,ds \\ + C_{11}\int_0^t \delta(s)\,ds + C_{12}\int_0^t\int_0^{h^-(s)} \big|u^1(x,s) - u^2(x,s)\big|\,dxds\end{aligned} \tag{4.8}$$

with suitable new constants C_{10} to C_{12}. To estimate the right-hand part of (4.8), let $z(x,s) := u^1(x,s) - u^2(x,s)$. This function satisfies in $Q(h^-,t)$ the linear equation

$$\partial_t z - \widetilde{a}\partial_x^2 z + \widetilde{b}\partial_x z + \widetilde{c}z = 0,$$

where the coefficients $\widetilde{a}$ to $\widetilde{c}$ are given by (the argument (x,s) is omitted wherever appropriate)

$$\widetilde{a} = a(u^1), \quad \widetilde{b} = a'(\partial_x u^1 + \partial_x u^2) + f'(u^1), \quad \widetilde{c} = \partial_x^2 u^2 \overline{a'} + (\partial_x u^2)^2 \overline{a''} + \partial_x u^2 \overline{f''},$$

where

$$\overline{g}(x,s) := \int_0^1 g\big(\lambda u^1(x,s) + (1-\lambda)u^2(x,s)\big)d\lambda, \quad g \in \Big\{a', a'', f', f'', \partial_2\widehat{\gamma}, \partial_2\widehat{\beta}\Big\}.$$

The function z satisfies the initial condition $z(x,0) = 0$ for $0 \le x \le 1$. From boundary condition (4.1c) and estimate (4.7) we obtain

$$\big((\overline{f'} - \partial_x u^2 \overline{a'})z - a\big(u^1\big)\partial_x z\big)(0,s) = \psi^1(s).$$

Similarly, boundary condition (4.1d) implies

$$\begin{aligned}\Big(\Big[\overline{f'} + \big[\widehat{\beta}\big(s,u^1\big)h^1(s) + \widehat{\gamma}\big(s,u^1\big)\big] + \overline{\partial_2\widehat{\beta}}h^2(s)u^2 \\ + \overline{\partial_2\widehat{\gamma}}u^2 + \overline{a'}(\partial_x u)^2\Big]z - a(u^1)\partial_x z\Big)\big(h^-(s),s\big) = \psi^2(s), \\ \psi^2(s) := -\widehat{\beta}\big(s,u^1(h^-(s),s)\big)\big(h^1(s) - h^2(s)\big)u^2\big(h^-(s),s\big).\end{aligned}$$

Since the functions $\widetilde{a}$ to $\widetilde{c}$ are bounded and since there exist constants C_{13} to C_{15} such that $|\widetilde{d}(x,s)| \le C_{13}\delta(t)$, $|\psi^1(s)| \le C_{14}\delta(s)$, and $|\psi^2(s)| \le C_{15}\delta(s)$, we obtain from the maximum principle that there exists a constant C_{16} independent of t with

$$\big|z(x,t)\big| \le C_{16}\max_{0\le s\le t}\delta(s);$$

hence inequality (4.8) reduces to

$$\delta(t) \le C_4\delta^2(t) + C_{17}\int_0^t \max_{0\le\tau\le s}\delta(\tau)\,ds.$$

Since $\delta(0) = 0$ and $\delta'(s)$ is uniformly bounded, we can choose a time $t_0 \in (0,T]$ such that $C_4\delta(t) \le 1/2$ for all $t \in (0,t_0]$. Thus

$$\delta(t) \le \frac{1}{2}\max_{0\le\tau\le t}\delta(\tau) + C_{17}\int_0^t \max_{0\le\tau\le s}\delta(\tau)\,ds \quad \text{for } 0 \le t \le t_0.$$

Consequently, there exists a constant C_{18} such that

$$\delta(t) \le C_{18} \int_0^t \max_{0\le\tau\le s} \delta(\tau)\, ds \quad \text{for } 0 \le t \le t_0.$$

This shows that $\delta(t) = 0$, i.e., $h^1(t) = h^2(t) =: h(t)$ for $0 \le t \le t_0$. The maximum principle then implies $u^1(x,t) = u^2(x,t)$ for $(x,t) \in Q(h,t_0)$, which contradicts the definition of t_1. Consequently, we obtain $u^1(x,t) = u^2(x,t)$ in $Q(h,T)$. This concludes the proof of Lemma 4.2. □

5. Existence of generalized solutions. To prove the existence of a generalized solution, we have to establish uniform estimates (with respect to the regularization parameter ε) of the solutions u_ε of the regularized free boundary problem (4.1). It is convenient to formulate these estimates in terms of the solutions $\{v_\varepsilon\}_{\varepsilon>0}$ of the problem (4.2) with fixed boundaries.

LEMMA 5.1. *Let $(v_\varepsilon, h_\varepsilon)$ be a solution of the regularized boundary problem* (4.2). *Then the following uniform estimates are valid, where the constant M_2 is independent of ε:*

$$\sup_{t\in[0,T]} \|\partial_x v_\varepsilon(\cdot,t)\|_{L^1(0,1)} \le M_2. \tag{5.1}$$

Proof. The proof closely follows that of Lemma 11 in [4]. Define approximations sgn_η and $|\cdot|_\eta$ of the sign and modulus functions by

$$\operatorname{sgn}_\eta(\tau) := \begin{cases} \operatorname{sgn}(\tau) & \text{if } |\tau| > \eta, \\ \tau/\eta & \text{if } |\tau| \le \eta, \end{cases} \quad |x|_\eta := \int_0^x \operatorname{sgn}_\eta(\zeta) d\zeta, \quad \eta > 0.$$

Setting $y_\varepsilon := \partial_\xi v_\varepsilon$, we obtain the following by differentiating (4.2a) with respect to ξ, multiplying it by $\operatorname{sgn}_\eta(y_\varepsilon)$, integrating over Q_{T_0}, where $0 < T_0 \le T$, and using integration by parts:

$$\begin{aligned} &\iint_{Q_{T_0}} \operatorname{sgn}_\eta(y_\varepsilon)\partial_t y_\varepsilon \, d\xi dt = \int_0^{T_0} \operatorname{sgn}_\eta(y_\varepsilon)\big(-\partial_\xi g_\varepsilon(v_\varepsilon,\xi,t) + h_\varepsilon^{-2}\partial_\xi^2 A_\varepsilon(v_\varepsilon)\big)\Big|_{\xi=0}^{\xi=1} dt \\ &\quad + \iint_{Q_{T_0}} \operatorname{sgn}'_\eta(y_\varepsilon)\partial_\xi y_\varepsilon \Big\{-h_\varepsilon^{-1}h'_\varepsilon\xi + h_\varepsilon^{-1}f'_\varepsilon(v_\varepsilon) - h_\varepsilon^{-2}a'_\varepsilon(v_\varepsilon)y_\varepsilon\Big\} y_\varepsilon \, d\xi dt \\ &\quad - \iint_{Q_{T_0}} \operatorname{sgn}'_\eta(y_\varepsilon) a_\varepsilon(v_\varepsilon)(\partial_\xi y_\varepsilon)^2 \, d\xi dt - \iint_{Q_{T_0}} \operatorname{sgn}_\eta(y_\varepsilon) h_\varepsilon^{-1} h'_\varepsilon y_\varepsilon \, d\xi dt \\ &=: I_\eta^1 + I_\eta^2 + I_\eta^3 + I_\eta^4. \end{aligned} \tag{5.2}$$

We now estimate the integrals I_η^1 to I_η^4. Using (4.2a), we see that

$$I_\eta^1 = \int_0^{T_0} \Big\{ \operatorname{sgn}_\eta\big(\partial_\xi v_\varepsilon(1,t)\big)\partial_t v_\varepsilon(1,t) - \operatorname{sgn}_\eta\big(\partial_\xi v_\varepsilon(0,t)\big)\partial_t v_\varepsilon(0,t)\Big\}\, dt.$$

The boundary conditions (4.2c) and (4.2d) imply that

$$\partial_\xi v_\varepsilon(0,t) = \frac{h_\varepsilon f_\varepsilon(v_\varepsilon(0,t))}{a_\varepsilon(v_\varepsilon(0,t))} \ge 0, \quad \partial_\xi v_\varepsilon(1,t) = \frac{h_\varepsilon[f_\varepsilon(v_\varepsilon(1,t)) - h'_\varepsilon v_\varepsilon(1,t)]}{a_\varepsilon(v_\varepsilon(1,t))} \ge 0. \tag{5.3}$$

In light of Lemma 4.1, we see from (5.3) that $\partial_\xi v_\varepsilon(0,t) = 0$ implies that $v_\varepsilon(0,t)$ assumes the constant value $v_{\varepsilon\min} := \inf \mathcal{U}^\varepsilon$ or $v_{\varepsilon\max} := \sup \mathcal{U}^\varepsilon$. Letting $\mathcal{E}_0 := \{t \in [0,T] : v_\varepsilon(0,t) = v_{\varepsilon\min} \text{ or } v_\varepsilon(0,t) = v_{\varepsilon\max}\}$, we see that $\partial_t v_\varepsilon(0,t) = 0$ a.e. in $\mathcal{E}_0$. We therefore conclude that

$$-\int_0^{T_0} \operatorname{sgn}_\eta\bigl(y_\varepsilon(0,t)\bigr)\partial_t v_\varepsilon(0,t)\,dt \overset{\eta\to 0}{\longrightarrow} -\int_0^{T_0} \partial_t v_\varepsilon(0,t)\,dt = v_\varepsilon(0,0) - v_\varepsilon(0,T_0).$$

Applying a similar argument to the boundary condition (4.2d), we obtain

$$I_\eta^1 \overset{\eta\to 0}{\longrightarrow} v_\varepsilon(1,T_0) - v_\varepsilon(1,0) + v_\varepsilon(0,0) - v_\varepsilon(0,T_0).$$

From Saks' lemma [2, 22] we infer that $I_\eta^2 \to 0$ for $\eta \to 0$. In light of $I_\eta^3 \le 0$ and

$$I_\eta^4 \overset{\eta\to 0}{\longrightarrow} -\iint_{Q_{T_0}} h_\varepsilon^{-1} h_\varepsilon' |y_\varepsilon|\,d\xi dt,$$

we get from (5.2)

$$\begin{aligned}\bigl\|\partial_x v_\varepsilon(\cdot,T_0)\bigr\|_{L^1(0,1)} \le{}& \|(u_\varepsilon^0)'\|_{L^1(0,1)} - v_\varepsilon(1,0) + v_\varepsilon(1,T_0) - v_\varepsilon(0,T_0)\\ &+ v_\varepsilon(0,0) + \int_0^{T_0} \bigl\|\partial_x v_\varepsilon(\cdot,t)\bigr\|_{L^1(0,1)}\,dt.\end{aligned}$$

An application of Gronwall's lemma yields estimate (5.1). ∎

For the present problem it is probably impossible to obtain a uniform $L^1(Q_T)$ estimate of the time derivative $\partial_t v_\varepsilon$, in contrast to several analyses of problems with fixed boundaries [4, 5]. For example, in [4] such an estimate was derived by differentiating the regularized parabolic equation with respect to t, multiplying the resulting equation by $\operatorname{sgn}_\eta(\partial_x v_\varepsilon)$, integrating the result over the computational domain, and using the boundary conditions and Gronwall's lemma. In the present case, differentiating (4.2a) with respect to t will produce an equation with a coefficient involving $h_\varepsilon''(t)$. However, we cannot bound this quantity, since differentiating the coupling equation (4.2f) with respect to t will lead to an equation for $h_\varepsilon''(t)$ in terms of $\partial_t v_\varepsilon$, and we cannot control the variation of v_ε with respect to t along the boundary $\xi = 0$.

To apply the compactness criterion to the family of regularized solutions $\{v_\varepsilon\}_{\varepsilon>0}$, we apply the following variant of Kružkov's [14] interpolation lemma (see, e.g., [13] for a proof).

LEMMA 5.2. *Assume that there exist finite constants c_1 and c_2 such that the function $u : (0,1)\times[0,T] \to \mathbb{R}$ satisfies $\|u(\cdot,t)\|_{L^\infty(0,1)} \le c_1$ and $\mathrm{TV}_{(0,1)}(u(\cdot,t)) \le c_2$ for all $t \in [0,T]$, and that u is weakly Lipschitz continuous with respect to t in the sense that*

$$\left|\int_0^1 \bigl(u(x,t_2) - u(x,t_1)\bigr)\varphi(x)\,dx\right| \le \mathcal{O}(t_2 - t_1)\sum_{i=0}^n \|\varphi^{(i)}\|_{L^\infty(0,1)}$$

for all $\varphi \in C_0^n(0,1)$, $0 \le t_1 \le t_2 \le T$. Then there exists a constant C, depending in particular on c_1 and c_2, such that the following interpolation result is valid:

$$\bigl\|u(\cdot,t_2) - u(\cdot,t_1)\bigr\|_{L^1(0,1)} \le C(t_2 - t_1)^{1/(n+1)}, \quad 0 \le t_1 \le t_2 \le T.$$

We calculate here that

$$\begin{aligned}
&\int_0^1 \big(v_\varepsilon(\xi,t_2)-v_\varepsilon(\xi,t_1)\big)\varphi(\xi)\,d\xi\\
&=\int_{t_1}^{t_2}\int_0^1 \Big\{h_\varepsilon^{-1}h_\varepsilon'(\xi\partial_\xi v_\varepsilon - v_\varepsilon) - h_\varepsilon^{-1}\partial_\xi f_\varepsilon(v_\varepsilon) + h_\varepsilon^{-2}\partial_\xi^2 A_\varepsilon(v_\varepsilon)\Big\}\varphi(\xi)\,d\xi dt\\
&=\int_{t_1}^{t_2}\int_0^1 \Big\{h_\varepsilon^{-1}h_\varepsilon' v_\varepsilon\varphi(\xi) + \big(-h_\varepsilon^{-1}h_\varepsilon'(t)\xi v_\varepsilon + h_\varepsilon^{-1}f_\varepsilon(v_\varepsilon) - h_\varepsilon^{-2}a_\varepsilon(v_\varepsilon)\partial_\xi v_\varepsilon\big)\varphi'(\xi)\Big\}\,d\xi dt.
\end{aligned}$$

From the proof of Lemma 4.4 it follows that there exists a constant $\widetilde{M}_h$ such that the estimate $\|1/h_\varepsilon^2\|_{L^\infty(0,T)} + \|h_\varepsilon'/h_\varepsilon\|_{L^\infty(0,T)} \le \widetilde{M}_h$ holds uniformly in ε. Using the estimate (5.1), we get

$$\begin{aligned}
&\left|\int_0^1 \big(v_\varepsilon(\xi,t_2)-v_\varepsilon(\xi,t_1)\big)\varphi(\xi)\,d\xi\right|\\
&\quad\le (t_2-t_1)\widetilde{M}_h\Big[M_0\|\varphi\|_{L^\infty(0,1)} + \big(\|f_\varepsilon\|_\infty + \|a_\varepsilon\|_\infty M_2 + M_0\big)\|\varphi'\|_{L^\infty(0,1)}\Big].
\end{aligned}$$

Thus we have proved the following.

LEMMA 5.3. *Let $(v_\varepsilon, h_\varepsilon)$ be a solution of the regularized boundary problem* (4.2). *Then the following uniform estimates are valid, where the constant M_3 is independent of ε:*

$$\big\|v_\varepsilon(\cdot,t_2)-v_\varepsilon(\cdot,t_1)\big\|_{L^1(0,1)} \le M_3(t_2-t_1)^{1/2}, \quad 0\le t_1\le t_2\le T. \tag{5.4}$$

In light of estimates (4.4), (5.1), and (5.4) of v_ε, a standard application of Kolmogorov's compactness criterion [23] yields that the family $\{v_\varepsilon\}$ is compact in $L^1(Q_T)$. Thus there exists a sequence $\varepsilon_n \to 0$ such that $\{v_{\varepsilon_n}\}$ converges in $L^1(Q_T)$ to a function $v \in BV_{1,1/2}(Q_T)$. Moreover, since the estimates of h_ε in (4.5) are uniform in ε, there exists a subsequence $\{h_{\varepsilon_n}\}$ of $\{h_\varepsilon\}$ and a function h such that $|h(t_2)-h(t_1)| \le M_h(t_2-t_1)$ for $0\le t_1\le t_2\le T$, $h(0)=1$ and h is nonincreasing.

We now have to prove that the limit pair (v,h) is indeed a generalized solution of the initial-boundary value problem (3.6). Obviously, the function h satisfies part (a) of Definition 3.2.

LEMMA 5.4. *The limit function v of solutions v_ε of the regularized problem* (4.2) *has the regularity properties stated in part* (b) *of Definition* 3.2.

Proof. Multiplying (4.2a) by v_ε and integrating the result over Q_T, we get

$$\begin{aligned}
\iint_{Q_T} h_\varepsilon^{-2}a_\varepsilon(v_\varepsilon)(\partial_\xi v_\varepsilon)^2 d\xi dt &= -\frac{1}{2}\int_0^1 v_\varepsilon^2\Big|_{t=0}^{t=T} d\xi - \int_{Q_T} h_\varepsilon^{-1}h_\varepsilon' v_\varepsilon^2\,d\xi dt\\
&\quad + \iint_{Q_T} g_\varepsilon(v_\varepsilon,\xi,t)\partial_\xi v_\varepsilon\,d\xi dt,
\end{aligned}$$

and thus

$$\big\|\partial_x A^\varepsilon(v_\varepsilon)\big\|_{L^2(Q_T)} \le \|a_\varepsilon\|_\infty\big\{M_0^2 + TM_h(2M_0^2 + M_2\|f_\varepsilon\|_\infty)\big\} =: M_4^\varepsilon.$$

The stated regularity of $A(u)$ follows by letting $\varepsilon\to 0$ and observing that M_4^ε is uniformly bounded for ε sufficiently small. To show the stated $\mathcal{DM}^2$ property, we

rewrite the regularized equation (4.2a) as follows, where $|k| \le K$ and K is a suitable large constant:

(5.5)
$$\partial_t(v_\varepsilon - k) + \partial_\xi\big(g_\varepsilon(v_\varepsilon,\xi,t) - g_\varepsilon(k,\xi,t)\big) + h_\varepsilon^{-1}h_\varepsilon'(v_\varepsilon - k) = h_\varepsilon^{-2}\partial_\xi^2\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big).$$

Multiplying (5.5) by $\operatorname{sgn}_\eta(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k))\zeta$, where $k \in \mathbb{R}$ and $\zeta \in C^\infty(\overline{Q_T})$ is an arbitrary test function, and integrating by parts over Q_T then yields

$$\begin{aligned}
&\iint_{Q_T} h_\varepsilon^{-2}\big[\partial_\xi\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\big]^2 \operatorname{sgn}_\eta'\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\zeta\, d\xi dt \\
&= \int_0^T \big(h_\varepsilon^{-2}\partial_\xi A_\varepsilon(v_\varepsilon) - g_\varepsilon(v_\varepsilon,\xi,t)\big)\operatorname{sgn}_\eta\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\zeta\Big|_{\xi=0}^{\xi=1} dt \\
&\quad + \int_0^T g_\varepsilon(k,\xi,t)\operatorname{sgn}_\eta\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\zeta\Big|_{\xi=0}^{\xi=1} dt \\
&\quad + \iint_{Q_T}\Big\{\big(g_\varepsilon(v_\varepsilon,\xi,t) - g_\varepsilon(k,\xi,t)\big) - h_\varepsilon^{-2}\partial_\xi\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\Big\} \\
&\qquad \times \operatorname{sgn}_\eta\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\partial_\xi\zeta\, d\xi dt \\
&\quad + \iint_{Q_T}\big(g_\varepsilon(v_\varepsilon,\xi,t) - g_\varepsilon(k,\xi,t)\big)\operatorname{sgn}_\eta'\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big) \\
&\qquad \times \partial_\xi\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\zeta\, d\xi dt \\
&\quad - \iint_{Q_T} h_\varepsilon^{-1}h_\varepsilon' v_\varepsilon \operatorname{sgn}_\eta\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\zeta\, d\xi dt + \int_0^1 |v_\varepsilon - k|_\eta\zeta\Big|_{t=0}^{t=T} d\xi \\
&\quad - \iint_{Q_T}(v_\varepsilon - k)\operatorname{sgn}_\eta'\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\partial_t\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\zeta d\xi dt \\
&\quad - \iint_{Q_T}(v_\varepsilon - k)\operatorname{sgn}_\eta\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\partial_t\zeta\, d\xi dt =: I_\eta^5 + \cdots + I_\eta^{12}.
\end{aligned} \tag{5.6}$$

We now consider the limit of the right-hand side of (5.6) for $\eta \to 0$. First note that $I_\eta^5 = 0$ due to the boundary conditions (4.2c) and (4.2d). By Lebesgue's theorem, we get

$$I_\eta^6 \xrightarrow{\eta\to 0} I_0^6 := \int_0^T g_\varepsilon(k,\xi,t)\operatorname{sgn}\big(A_\varepsilon(v_\varepsilon(\xi,t)) - A_\varepsilon(k)\big)\zeta(\xi,t)\Big|_{\xi=0}^{\xi=1} dt,$$

which implies $|I_0^6| \le T\|g_\varepsilon\|_\infty\|\zeta\|_\infty$. Using the properties of sgn_η, Lebesgue's theorem, $\partial_\xi A(k) = 0$, and the fact that $\operatorname{sgn}(v_\varepsilon - k) = \operatorname{sgn}(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k))$ due the monotonicity of $A_\varepsilon(\cdot)$, we get

$$I_\eta^7 \xrightarrow{\eta\to 0} \iint_{Q_T}\Big\{\operatorname{sgn}(v_\varepsilon - k)\big(g_\varepsilon(v_\varepsilon,\xi,t) - g_\varepsilon(k,\xi,t)\big) - h_\varepsilon^{-2}\partial_\xi\big|A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big|\Big\}\,\partial_\xi\zeta d\xi dt.$$

Precisely as in [18], using that $u\operatorname{sgn}_\eta'(u) \le \chi_{\{u:\,0<|u|\le\eta\}}$ and recalling from assumption (4.3) that the inverse function A_ε^{-1} is for fixed ε Lipschitz continuous with constant $1/\nu_\varepsilon$, we get that

$$\begin{aligned}
&\big|\big(g_\varepsilon(v_\varepsilon,\xi,t) - g_\varepsilon(k,\xi,t)\big)\operatorname{sgn}_\eta'\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\partial_\xi\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\big| \\
&\qquad\qquad \le \frac{L_\varepsilon}{\nu_\varepsilon}\big|\partial_\xi\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\big|\chi_{\mathcal{I}(\varepsilon,\eta)},
\end{aligned}$$

where $\mathcal{I}(\varepsilon,\eta) := \{(\xi,t):\ 0 \le |A_\varepsilon(v_\varepsilon(\xi,t)) - A_\varepsilon(k)| \le \eta\}$. Consequently,

$$|I_\eta^8| \le \frac{L_\varepsilon}{\nu_\varepsilon}\|\zeta\|_{L^\infty(Q_T)} \iint_{\mathcal{I}(\varepsilon,\eta)} \big|\partial_\xi\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\big|\, d\xi dt.$$

Observe that $\operatorname{meas}\mathcal{I}(\varepsilon,\eta) \to 0$ as $\eta \to 0$, since this measure converges to that of the empty set. Thus $I_\eta^8 \to 0$ as $\eta \to 0$. Next, we see that

$$I_\eta^9 \overset{\eta\to 0}{\longrightarrow} I_0^9 := -\iint_{Q_T} h_\varepsilon^{-1} h_\varepsilon'(t) v_\varepsilon \operatorname{sgn}\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\zeta\, d\xi dt$$

with $|I_0^9| \le T M_h M_0 \|\zeta\|_{L^\infty(Q_T)}$. Furthermore,

$$I_\eta^{10} \overset{\eta\to 0}{\longrightarrow} I_0^{10} := \int_0^1 \Big\{\big|v_\varepsilon(\xi,T) - k\big|\zeta(\xi,T) - \big|u_0^\varepsilon(\xi) - k\big|\zeta(\xi,0)\Big\} d\xi,$$

and thus $|I_0^{10}| \le 2(M_0 + K)\|\zeta\|_{L^\infty(Q_T)}$. The integrand of I_η^{11} satisfies

$$\begin{aligned}
&\big|(v_\varepsilon - k)\operatorname{sgn}_\eta'\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\partial_t\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\zeta\big| \\
&= \big|(v_\varepsilon - k)\operatorname{sgn}_\eta'(v_\varepsilon - k)\partial_t\big(A_\varepsilon(u_\varepsilon) - A_\varepsilon(k)\big)\zeta\big| \\
&\le \big|\partial_t\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\big|\chi_{\{(\xi,t):\, 0\le|v_\varepsilon(\xi,t)-k|\le\eta\}}.
\end{aligned}$$

An argument similar to that employed for I_η^8 reveals that $I_\eta^{11} \to 0$ as $\eta \to 0$. Finally, we obtain

$$I_\eta^{12} \overset{\eta\to 0}{\longrightarrow} I_0^{12} := -\iint_{Q_T} |v_\varepsilon - k|\partial_\xi\zeta\, d\xi dt.$$

Collecting the estimates of I_η^5 to I_η^{12} yields that all terms of the right-hand part of (5.6) possess a limit as $\eta \to 0$ and are in particular uniformly bounded with respect to η. Thus, taking $\zeta \equiv 1$ we see that there exists a constant C_1, depending possibly on ε (but not on η), such that

$$\iint_{Q_T} h_\varepsilon^{-2}\big[\partial_\xi\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\big]^2 \operatorname{sgn}_\eta'\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\, d\xi dt \le C_1(\varepsilon).$$

Consequently, the sequence

$$\{E_{\varepsilon,\eta}\}_{\eta>0} := \Big\{\big(h_\varepsilon(t)\big)^{-2}\big[\partial_\xi\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\big]^2 \operatorname{sgn}_\eta'\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\Big\}_{\eta>0}$$

is bounded in $L^1(Q_T)$ with respect to η and therefore also in $C(\overline{Q_T})'$, the dual of the space $C(\overline{Q_T})$ of continuous functions on $\overline{Q_T}$. By compactness of the weak-$\star$ topology of $C(\overline{Q_T})'$ we deduce that, up to subsequences, the sequence $\{E_{\varepsilon,\eta}\}_\eta$ converges towards an element $E_\varepsilon \in C(\overline{Q_T})'$ in the weak-$\star$ topology. Thus for any $\zeta \in C^\infty(\overline{Q_T})$

we can pass to the limit $\eta \to 0$ in (5.6) to obtain

$$
\begin{aligned}
\langle E_\varepsilon, \zeta\rangle = &\int_0^T g_\varepsilon(k,\xi,t)\,\mathrm{sgn}\big(A_\varepsilon(v_\varepsilon(\xi,t)) - A_\varepsilon(k)\big)\zeta(\xi,t)\Big|_{\xi=0}^{\xi=1} dt \\
&- \iint_{Q_T} h_\varepsilon^{-1} h_\varepsilon'(t) v_\varepsilon\,\mathrm{sgn}\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\zeta\,d\xi dt \\
&+ \iint_{Q_T} \Big\{\mathrm{sgn}(v_\varepsilon - k)\big(g_\varepsilon(v_\varepsilon,\xi,t) - g_\varepsilon(k,\xi,t)\big) \\
&\qquad - h_\varepsilon^{-2}\partial_\xi\big|A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big|\Big\}\partial_\xi\zeta\,d\xi dt \\
&+ \int_0^1 \Big\{\big|v_\varepsilon(\xi,T) - k\big|\zeta(\xi,T) - \big|u_0^\varepsilon(\xi) - k\big|\zeta(\xi,0)\Big\}d\xi \\
&- \iint_{Q_T} |v_\varepsilon - k|\partial_t\zeta\,d\xi dt.
\end{aligned}
\tag{5.7}
$$

On the other hand, due to the properties of the function sgn_η, we have $E_{\varepsilon,\eta} \geq 0$ for every $\varepsilon, \eta > 0$. Therefore we get

$$
\begin{aligned}
&\frac{|\langle E_\varepsilon, \zeta\rangle|}{\|\zeta\|_{L^\infty(Q_T)}} \\
&= \lim_{\eta\to 0} \frac{1}{\|\zeta\|_{L^\infty(Q_T)}} \left| \iint_{Q_T} h_\varepsilon^{-2}\big[\partial_\xi\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\big]^2\,\mathrm{sgn}_\eta'\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\zeta\,d\xi dt \right| \\
&\leq \limsup_{\eta\to 0} \iint_{Q_T} h_\varepsilon^{-2}\big[\partial_\xi\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\big]^2\,\mathrm{sgn}_\eta'\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)d\xi dt.
\end{aligned}
$$

Thus we get from (5.7) with $\zeta \equiv 1$

$$
\begin{aligned}
\frac{|\langle E_\varepsilon, \zeta\rangle|}{\|\zeta\|_{L^\infty(Q_T)}} \leq &- \iint_{Q_T} h_\varepsilon^{-1} h_\varepsilon'(t) v_\varepsilon\,\mathrm{sgn}\big(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)\big)\,d\xi dt \\
&+ \int_0^T g_\varepsilon(k,\xi,t)\,\mathrm{sgn}\big(A_\varepsilon(v_\varepsilon(\xi,t)) - A_\varepsilon(k)\big)\Big|_{\xi=0}^{\xi=1} dt \\
&+ \int_0^1 \Big\{\big|v_\varepsilon(\xi,T) - k\big| - \big|u_0^\varepsilon(\xi) - k\big|\Big\}d\xi.
\end{aligned}
\tag{5.8}
$$

Using the estimate (4.4) we deduce that there exists a constant C_2, which does not depend on ε, such that $|\langle E_\varepsilon, \zeta\rangle| \leq C_2\|\zeta\|_{L^\infty(Q_T)}$ for all $\varepsilon > 0$. Consequently, E_ε is bounded in $C(\overline{Q_T})'$, and up to a subsequence E_ε converges in the weak-$\star$ topology to a functional $E \in C(\overline{Q_T})'$, i.e., a Radon measure. We now pass to the limit $\varepsilon \to 0$ in (5.7). Since $g_\varepsilon(k,\xi,t)$ converges strongly to $g(k,\xi,t)$ and $\mathrm{sgn}(A_\varepsilon(v_\varepsilon) - A_\varepsilon(k))$ is bounded, $g_\varepsilon(k,\xi,t)\,\mathrm{sgn}(A_\varepsilon(v_\varepsilon(\xi,t)) - A_\varepsilon(k))$ converges weakly in $L^1(\{\xi\}\times(0,T))$, where $\xi = 0$ or $\xi = 1$, to $g(k,\xi,t)\,\mathrm{sgn}(A(v(\xi,t)) - A(k))$. Moreover, $|v_\varepsilon - k|$ converges strongly to $|v-k|$ in $C(0,T;L^1(0,1))$, $g_\varepsilon(v_\varepsilon,\xi,t)$ converges strongly to $g(v,\xi,t)$ in $L^q(Q_T)$ for every $q < \infty$, and $\partial_x|A_\varepsilon(v_\varepsilon) - A_\varepsilon(k)|$ converges weakly in $L^2(Q_T)$ to $\partial_\xi|A(v) - A(k)|$. Passing to the limit $\varepsilon \to 0$ in (5.7) we conclude that for all $\varphi \in C_0^\infty(Q_T)$,

$$
\begin{aligned}
\langle E, \varphi\rangle = &- \iint_{Q_T} h^{-1}h' v\,\mathrm{sgn}\big(A(v) - A(k)\big)\varphi\,d\xi dt - \iint_{Q_T} |v-k|\partial_t\varphi\,d\xi dt \\
&+ \iint_{Q_T} \Big\{\mathrm{sgn}(v-k)\big(g(v,\xi,t) - g(k,\xi,t)\big) - h^{-2}\partial_\xi\big|A(v) - A(k)\big|\Big\}\partial_\xi\varphi\,d\xi dt.
\end{aligned}
\tag{5.9}
$$

Since g, $\operatorname{sgn}(A(v)-A(k))$, and $\partial_\xi|A(v)-A(k)|$ are all functions in $L^1(Q_T)$, and since E is a Radon measure, we obtain from (5.9) that for all $\varphi \in C_0^\infty(Q_T)$,

$$\left| \iint_{Q_T} \Big\{ |v-k|\partial_t\varphi + \Big(\operatorname{sgn}(v-k)\big(g(v,\xi,t)-g(k,\xi,t)\big) - h^{-2}\partial_\xi\big|A(v)-A(k)\big|\Big)\partial_\xi\varphi \Big\}\, d\xi dt \right| \le C\|\varphi\|_{L^\infty(Q_T)}.$$

This in particular implies the stated $\mathcal{DM}^2$ property (3.9). ∎

LEMMA 5.5. *The limit function v of solutions v_ε of the regularized initial-boundary value problem satisfies the boundary conditions* (3.3) *and* (3.4) *stated in Definition* 3.1.

Proof. First of all we have from Lemma 4.2, passing to a subsequence if necessary, that h_ε converges uniformly to a certain Lipschitz function h, which satisfies $h(0)=1$, $h(t)\ge h_0>0$. Multiplying (4.1a) by $\varphi \in C_0^1(\Pi_T)$, integrating over $Q(h_\varepsilon,T)$, using integration by parts and the boundary conditions (4.1c), (4.1d), and then letting $\varepsilon \to 0$, we get

$$\iint_{Q(h,T)} \Big\{ u\partial_t\varphi + \big(f(u)-\partial_x A(u)\big)\partial_x\varphi \Big\}\, dx\, dt = 0. \tag{5.10}$$

From (5.10) there follow two conclusions about the $\mathcal{DM}^2$ field $F=(F_1,F_2)=(f(u)-\partial_x A(u),u)$: $\operatorname{div} F = 0$ (this is the obvious one) and $\langle F\cdot\nu|_{\partial Q(h,T)},\varphi\rangle = 0$, as a consequence of the generalized Gauss–Green formula (2.7). Hence, using (2.8) and (2.9) we deduce (3.3) and (3.4). ∎

LEMMA 5.6. *The limit function (u,h) of solutions $(u_\varepsilon,h_\varepsilon)$ of the regularized problem* (4.1) *satisfies* (2.5e) *in the sense stated in* (d) *of Definition* 3.1.

Proof. First, we observe that $A_\varepsilon(u_\varepsilon(x,t))$ converges to $A(u(x,t))$ in $L^1_{\mathrm{loc}}(Q(h,T))$. This follows by the convergence of $A_\varepsilon(v_\varepsilon(\xi,t))$ to $A(v(\xi,t))$ in $L^1(Q_T)$, the uniform convergence of h_ε to h, and the uniform boundedness of $\partial_\xi A_\varepsilon(v_\varepsilon(\xi,t))$ in $L^2(Q_T)$. More specifically, for any compact $K\subset Q(h,T)$, for ε sufficiently small,

$$\begin{aligned}
&\iint_K \big|A_\varepsilon\big(u_\varepsilon(x,t)\big)-A\big(u(x,t)\big)\big|\,dx\,dt\\
&= \iint_{K'} \big|A_\varepsilon\big(u_\varepsilon(h(t)\xi,t)\big)-A\big(u(h(t)\xi,t)\big)\big|h(t)\,d\xi\,dt\\
&\le \iint_{K'} \Big\{ \big|A_\varepsilon\big(v_\varepsilon(\xi,t)\big)-A\big(v(\xi,t)\big)\big| + \big|A_\varepsilon\big(u_\varepsilon(h(t)\xi,t)\big)-A_\varepsilon\big(u_\varepsilon(h_\varepsilon(t)\xi,t)\big)\big| \Big\} h(t)\,d\xi dt\\
&\le \iint_{K'} \big|A_\varepsilon\big(v_\varepsilon(\xi,t)\big)-A\big(v(\xi,t)\big)\big|h(t)\,d\xi\,dt\\
&\quad + C\|h_\varepsilon-h\|_\infty \sup_\varepsilon \|\partial_x A_\varepsilon(u_\varepsilon)\|_{L^2(Q(h_\varepsilon,T))} \overset{\varepsilon\to0}{\longrightarrow} 0,
\end{aligned}$$

where K' denotes the image of K by the transformation $(x,t)\mapsto(\xi,t)$. Now, we prove that $A_\varepsilon(u_\varepsilon(h_\varepsilon(t),t)) \to \gamma_{x\to h(t)}A(u(\cdot,t))$ in $L^1(0,T)$ as $\varepsilon\to0$, after passing to a suitable subsequence if necessary. Given any $\delta>0$, we have $h(t)-\delta<h_\varepsilon(t)<h(t)+\delta$, $0<t<T$, for ε sufficiently small, due to the uniform convergence $h_\varepsilon\to h$. We may also assume that $A_\varepsilon(u_\varepsilon(h(t)-\delta,t))\to A(u(h(t)-\delta,t))$ in $L^1(0,T)$ due to the convergence of $A_\varepsilon(u_\varepsilon(x,t))$ to $A(u(x,t))$ in $L^1_{\mathrm{loc}}(Q(h,T))$. Then, setting

$B_\varepsilon(x,t) := A_\varepsilon(u_\varepsilon(x,t))$, $B(x,t) := A(u(x,t))$, and $x_\delta(t) := h(t) - \delta$, we have

$$\begin{aligned}\int_0^T \big|B_\varepsilon\big(h_\varepsilon(t),t\big) - \gamma_{x\to h(t)}B(\cdot,t)\big|\,dt &\le \int_0^T \big|B_\varepsilon\big(x_\delta(t),t\big) - B\big(x_\delta(t),t\big)\big|\,dt \\ &\quad + \int_0^T \big|B_\varepsilon\big(x_\delta(t),t\big) - B_\varepsilon\big(h_\varepsilon(t),t\big)\big|\,dt + \int_0^T \big|B\big(x_\delta(t),t\big) - \gamma_{x\to h(t)}B(\cdot,t)\big|\,dt \\ &\le \int_0^T \big|B_\varepsilon\big(x_\delta(t),t\big) - B\big(x_\delta(t),t\big)\big|\,dt + C\sqrt{\delta}.\end{aligned}$$

Since $\delta > 0$ may be taken arbitrarily small, the assertion follows. Finally, by passing to a further subsequence of ε's if necessary, we see that, except for $h'_\varepsilon(t)$, all other terms in (4.1e) converge a.e. in $(0,T)$ to the corresponding terms in (2.5e), replacing $A(u(h(t),t))$ by $\gamma_{x\to h(t)}A(u(\cdot,t))$. Therefore, $h'_\varepsilon(t)$ also converges a.e. in $(0,T)$, and since it clearly converges weakly to $h'(t)$, we have $h'_\varepsilon(t) \to h(t)$ a.e. in $(0,T)$, and the lemma is proved. □

It is standard to conclude from Lemma 5.2 that the limit function v satisfies the initial condition (3.12), and to prove that the entropy inequality (3.13) is satisfied by multiplying (4.2a) by $\operatorname{sgn}_\eta(v_\varepsilon - k)\varphi$, $k \in \mathbb{R}$, $\varphi \in C_0^\infty(Q_T)$, $\varphi \ge 0$, and letting $\eta \to 0$ and $\varepsilon \to 0$. Thus we have shown the following.

THEOREM 5.7. *The initial-boundary value problem* (3.6) *admits an entropy solution* (v,h).

Since $h(t) > 0$ and h' is bounded, we conclude that the following holds.

COROLLARY 5.8. *The free boundary problem* (2.5) *admits an entropy solution* (u,h).

REFERENCES

[1] G. ANZELLOTTI, *Pairings between measures and bounded functions and compensated compactness*, Ann. Mat. Pura Appl. (4), 135 (1983), pp. 293–318.

[2] C. BARDOS, A.Y. LE ROUX, AND J.C. NEDELEC, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[3] R. BÜRGER, F. CONCHA, AND K.H. KARLSEN, *Phenomenological model of filtration processes: 1. Cake formation and expression*, Chem. Engrg. Sci., 56 (2001), pp. 4537–4553.

[4] R. BÜRGER, S. EVJE, AND K.H. KARLSEN, *On strongly degenerate convection-diffusion problems modeling sedimentation-consolidation processes*, J. Math. Anal. Appl., 247 (2000), pp. 517–556.

[5] R. BÜRGER AND W.L. WENDLAND, *Existence, uniqueness, and stability of generalized solutions of an initial-boundary value problem for a degenerating quasilinear parabolic equation*, J. Math. Anal. Appl., 218 (1998), pp. 207–239.

[6] R. BÜRGER AND W.L. WENDLAND, *Entropy boundary and jump conditions in the theory of sedimentation with compression*, Math. Methods Appl. Sci., 21 (1998), pp. 865–882.

[7] R. BÜRGER, W.L. WENDLAND, AND F. CONCHA, *Model equations for gravitational sedimentation-consolidation processes*, ZAMM Z. Angew. Math. Mech., 80 (2000), pp. 79–92.

[8] M.C. BUSTOS, F. CONCHA, R. BÜRGER, AND E.M. TORY, *Sedimentation and Thickening: Phenomenological Foundation and Mathematical Theory*, Kluwer Academic, Dordrecht, The Netherlands, 1999.

[9] G.-Q. CHEN AND H. FRID, *Divergence-measure fields and hyperbolic conservation laws*, Arch. Ration. Mech. Anal., 147 (1999), pp. 89–118.

[10] G.-Q. CHEN AND H. FRID, *The theory of divergence-measure fields and applications*, Bol. Soc. Brasil. Mat. (N.S.), 32 (2001), pp. 401–433.

[11] L.C. EVANS AND R.F. GARIEPY, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[12] P. GARRIDO, F. CONCHA, AND R. BÜRGER, *Application of the unified model of solid-liquid separation of flocculated suspensions to experimental results*, in Proceedings of the Sixth Southern Hemisphere Meeting on Mineral Technology, Rio de Janeiro, Brazil, 2001, Vol. 1,

A.B. da Luz, P.S.M. Soares, M.L. Torem, and R.B.E. Trinidade, eds., CETEM/MCT, Rio de Janeiro, Brazil, 2001, pp. 117–122.

[13] K.H. Karlsen and N.H. Risebro, *Corrected operator splitting for nonlinear parabolic equations*, SIAM J. Numer. Anal., 37 (2000), pp. 980–1003.

[14] S.N. Kružkov, *First order quasilinear equations in several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.

[15] O.A. Ladyženskaja, V.A. Solonnikov, and N.N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[16] K.A. Landman and W.B. Russel, *Filtration at large pressures for strongly flocculated suspensions*, Phys. Fluids A, 5 (1993), pp. 550–560.

[17] J. Málek, J. Nečas, M. Rokyta, and M. Ružička, *Weak and Measure-Valued Solutions to Evolutionary PDEs*, Chapman & Hall, London, 1996.

[18] C. Mascia, A. Porretta, and A. Terracina, *Non-homogeneous Dirichlet problems for degenerate parabolic-hyperbolic equations*, Arch. Ration. Mech. Anal., 163 (2002), pp. 87–124.

[19] F. Otto, *First Order Equations with Boundary Conditions*, Preprint, SFB 256, University of Bonn, Germany, 1992.

[20] F. Otto, *Ein Randwertproblem für Erhaltungssätze*, Doctoral thesis, University of Bonn, Germany, 1993.

[21] F. Otto, *Initial-boundary value problem for a scalar conservation law*, C.R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 729–734.

[22] S. Saks, *Theory of the Integral*, Monografie Matematyczne, Drukarna Uniw. Jagiellonskiego, Warsaw, Poland, G.E. Stechert, New York, 1937.

[23] J. Wloka, *Partielle Differentialgleichungen*, Teubner Verlag, Stuttgart, 1982.

[24] Z. Wu, *A note on the first boundary value problem for quasilinear degenerate parabolic equations*, Acta Math. Sci. (English Ed.), 4 (1982), pp. 361–373.

[25] Z. Wu, *A Boundary Value Problem for Quasilinear Degenerate Parabolic Equations*, MRC Technical Summary Report #2484, University of Wisconsin, Madison, WI, 1983.

[26] Z. Wu and J. Wang, *Some results on quasilinear degenerate parabolic equations of second order*, in Proceedings of the 1980 Beijing Symposium on Differential Geometry and Differential Equations, Vol. 3, Science Press, Beijing, Gordon and Breach, New York, 1982, pp. 1593–1609.

[27] Z. Wu, J. Zhao, J. Yin, and H. Li, *Nonlinear Diffusion Equations*, World Scientific, Singapore, 2001.

[28] J. Zhao and Y. Li, *A free boundary problem for quasilinear degenerate parabolic equations with general degeneracy*, Acta Math. Sinica (N.S.), 6 (1990), pp. 364–382.

SIAM J. MATH. ANAL.
Vol. 34, No. 3, pp. 636–674

A GENERAL FRAMEWORK FOR DIFFRACTIVE OPTICS AND ITS APPLICATIONS TO LASERS WITH LARGE SPECTRUMS AND SHORT PULSES*

KAREN BARRAILH† AND DAVID LANNES‡

Abstract. The aim of this article is to generalize the usual tools of diffractive optics in order to allow the study of phenomena which are out of their range. This generalization relies on the algebra of oscillations with a continuous oscillatory spectrum, which is wider than the usual spaces of periodic or almost-periodic functions. We perform the analysis for general nonlinear hyperbolic systems, both in the dispersive and in the nondispersive cases, and particularly focus on the behavior of the nonlinearities. Our tools yield considerable simplifications in these nonlinearities, which allows us to point out qualitative differences between the dispersive and the nondispersive cases. Finally, we study in detail two physical examples which can be modeled with the present tools: lasers with large spectrums, and those with ultrashort pulses.

Key words. nonlinear hyperbolic systems, Maxwell equations, diffractive optics, continuous oscillatory spectrum, large spectrum, short pulse, nonlinear Schrödinger equation

AMS subject classifications. 28B05, 35L40, 35Q55, 35Q60, 35S99

PII. S0036141001398976

1. General comment.

1.1. Introduction.

Maxwell equations and many of the physical systems encountered in optics may be written in the form

$$\begin{cases} L^\varepsilon(\partial)\mathbf{u}^\varepsilon + f(\mathbf{u}^\varepsilon) = 0, \\ \mathbf{u}^\varepsilon|_{T=0}(X,Y,Z) = \mathbf{u}^0_\varepsilon(X,Y,Z), \end{cases} \tag{1.1}$$

where $\mathbf{u}^\varepsilon$ takes its values in R^n, and $L^\varepsilon(\partial)$ is a hyperbolic symmetric operator which one writes as

$$L^\varepsilon(\partial) = A_0\partial_T + A_1\partial_X + A_2\partial_Y + A_3\partial_Z + \frac{L_0}{\varepsilon},$$

the matrices A_i being symmetric and L_0 skew-symmetric.

In the study of the propagation of a diffractive laser beam with frequency ω_l and wavenumber $\overrightarrow{k_l} = (0,0,k_l)$, an approximate solution u^ε of $\mathbf{u}^\varepsilon$ is generally sought in the form

$$u^\varepsilon(T,X,Y,Z) = \varepsilon^p(\mathcal{U}_0(\varepsilon T,T,X,Y,Z)e^{i(\omega_l T - k_l Z)/\varepsilon} + \text{c.c.}), \tag{1.2}$$

the exponent p being chosen in order for the nonlinear and diffractive effects to come into play at the same time scale. The method of diffractive optics (see [8], for instance) consists of finding some equations which determine the *profile* $\mathcal{U}_0$.

The object of this article is to introduce a general framework for diffractive optics, which generalizes classical studies in both dispersive and nondispersive diffractive

*Received by the editors November 29, 2001; accepted for publication (in revised form) May 8, 2002; published electronically January 7, 2003.
http://www.siam.org/journals/sima/34-3/39897.html
†MAB, Université Bordeaux 1, 351 Cours de la Libération, 33405 Talence Cedex, France, and CEA Cesta, BP 233114 Le Barp, France (barrailh@bordeaux.cea.fr).
‡MAB, Université Bordeaux 1 et CNRS UMR 5466, 351 Cours de la Libération, 33405 Talence Cedex, France (lannes@math.u-bordeaux.fr).

optics, as shown in [10] and [11], for instance, and allows us also, without added difficulty, to treat more pathological situations. Among these, we study here two physical problems which cannot be modeled by oscillations of type (1.2). These physical phenomena are *large-spectrum lasers* and *ultrashort pulses*.

Large spectrum. The oscillatory spectrum of a classical oscillation of type (1.2) is located at two points, $\{\pm(\omega_l, -\overrightarrow{k_l})\}$. Experimentally, such a localization for the spectrum is never realized. Physically, the spectrum is concentrated around $\{\pm(\omega_l, -\overrightarrow{k_l})\}$, but it never reduces to these two points. These variations are generally taken into account in the amplitude $\mathcal{U}_0$. However, modifying $\mathcal{U}_0$ can only make the spectrum of u^ε expand around $\{\pm(\omega_l, -\overrightarrow{k_l})\}$ in an $O(\varepsilon)$ range. Lasers with large spectrum typically have a spectrum of width $O(1)$ and therefore cannot be modeled with usual oscillations of type (1.2). Direct computations for lasers with large spectrums have been carried out by Morice [13]. Here, we choose another approach and seek an approximate solution of (1.1) in the form

$$u^\varepsilon(T,X,Y,Z) = \varepsilon^p\big(\mathcal{U}_{0,I,1}(\varepsilon T,T,X,Y,Z)e^{i(\omega_l T - k_l Z)/\varepsilon} + \text{c.c.}\big) + \varepsilon^p\,\mathcal{U}_{0,II}\left(\varepsilon T,T,X,Y,Z,\frac{T}{\varepsilon},\frac{Z}{\varepsilon}\right),$$

where $\mathcal{U}_{0,II}$ is an oscillation with a *purely continuous spectrum*. We introduce this notion in Proposition 1.10 below; for the moment, just assume that $\mathcal{U}_{0,II}$ is smooth and decaying in its last two variables.

Considering a model of nonlinear Maxwell equations (system (M) in section 4.1), we prove that the amplitude $\mathcal{E}_{0,I,1}$ of the oscillating component of the electric field satisfies the usual nonlinear Schrödinger (NLS) equation

$$\partial_\tau \mathcal{E}_{0,I,1} + i\frac{\omega'(k_l)}{2k_l}(\partial_X^2+\partial_Y^2)\mathcal{E}_{0,I,1} + i\frac{\omega''(k_l)}{2}\partial_Z^2\mathcal{E}_{0,I,1} = \text{Cst}\ |\mathcal{E}_{0,I,1}|^2\mathcal{E}_{0,I,1},$$

while the corrective term $\mathcal{E}_{0,II}$ satisfies a linear equation,

$$\partial_\tau\partial_{z_0}\mathcal{E}_{0,II} - \frac{\omega'(D_{z_0})}{2}(\partial_X^2+\partial_Y^2)\mathcal{E}_{0,II} - \frac{D_{z_0}\omega''(D_{z_0})}{2}\partial_Z^2\mathcal{E}_{0,II} = 0.$$

The main interest of this latter equation is that all the nonlinearities one would find by a direct computation have been dropped, making this equation linear when it was a priori nonlinear. This fact is a striking consequence of the general results proved thereafter.

Ultrashort pulses. Seeking an approximate solution in the form (1.2) supposes that the profile $\mathcal{U}_0$ varies little compared with the scale of an oscillation. This condition is satisfied for almost all lasers because the length of the pulse is great compared with the wavelength. For the ultrashort pulses, obtained by recent lasers, this is no longer the case (see Figure 1). We refer to [4] and the references therein for a brief history of the study of short pulses in geometric optics. For diffractive time scales, the most general studies we know have been performed by Alterman and Rauch; see [1], [2], [3], and [15]. In the *nondispersive* case, these authors proved rigorously, using asymptotic techniques, that the Schrödinger approximation used for wave trains must be replaced by another approximation for short pulses,

$$2\partial_{z_0}\partial_\tau\mathcal{V} = v(\partial_X^2+\partial_Y^2)\mathcal{V} + \partial_{z_0}f(\mathcal{V}), \tag{1.3}$$

where $v\overrightarrow{e}_Z$ denotes the group velocity.

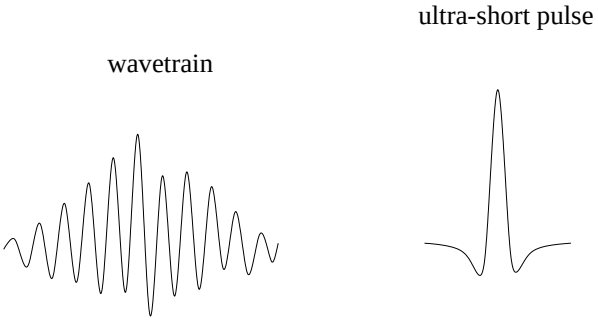


FIG. 1. *Example of a wave train and a short pulse.*

One of the main interests of [1], [3], and [15] is that these papers address pulses which may not have vanishing mean, as was the case in earlier papers [16], [17]. We use here Alterman's technique of infrared cutoffs to obtain this generality, but otherwise our approach is completely different since it is based on oscillations with continuous spectrum. We see three main interests in our method. First, it generalizes the usual methods of diffractive optics [8], [10], [11], so that short pulses do not appear as a pathological case, and "mixed" cases, such as the above lasers with large spectrums, can be addressed without added difficulty. The second interest resides in the study of the nonlinearities, since we prove that most of them can be dropped because their influence is negligible. Finally, we are able to address dispersive models, which are physically the most relevant.

More precisely, an approximate solution for the short pulse is sought in the form

$$u^\varepsilon(T,X,Y,Z) = \varepsilon^p \mathcal{U}_{0,II}\left(\varepsilon T, T, X, Y, Z, \frac{T}{\varepsilon}, \frac{Z}{\varepsilon}\right),$$

where $\mathcal{U}_{0,II}$ is a profile with a purely continuous oscillatory spectrum. Indeed, the pulse here is too short for a sinusoidal oscillation to appear. In the nondispersive case, we find of course that $\mathcal{U}_{0,II}$ must satisfy Alterman and Rauch's equation (1.3). The dispersive case is both simpler and more complicated because if the nonlinearities of (1.3) can be neglected, the group velocity $v\overrightarrow{e}_Z$ then depends on the frequency.

Remark 1.1. As the common formalism suggests, short pulses and large spectrum corrections to wave trains are essentially the same. The former focus on the time domain, while the latter look at the Fourier domain.

1.2. The spaces. We seek approximate solutions of (1.1) for diffractive time scales. Therefore, three scales of variables are used in this study:

- the fast scale $O(\frac{1}{\varepsilon})$ of the oscillations,
- the intermediate scale $O(1)$ of geometrical optics,
- the slow scale $O(\varepsilon)$ related to diffractive effects.

In order to identify clearly the variations of the solutions in these scales, auxiliary functions named *profiles* are introduced as in [8], and we look for exact solutions of (1.1) in the form

$$\mathbf{u}^\varepsilon = \varepsilon^p \mathbf{U}^\varepsilon\left(\varepsilon T, T, X, Y, Z, \frac{T}{\varepsilon}, \frac{Z}{\varepsilon}\right).$$

The factor ε^p is chosen to have both nonlinear and diffractive effects on the same time scale.

Before introducing the spaces associated to the profiles $\mathcal{V}(\tau, T, X, Y, Z, t_0, z_0)$ that we use to represent the exact and approximate solutions of (1.1), let us set some notation.

Notation. From now on, we write $\theta := (\omega_l t_0 - k_l z_0)$ and denote by $\xi := (\omega, k)$ and $\eta := (\eta_1, \eta_2, \eta_3)$ the Fourier dual variable of (t_0, z_0) and (X, Y, Z), respectively. The letter s will always denote a positive real number $s > 3/2$.

We also denote by $\mathcal{F}$ and by $\widehat{\ }$ the Fourier transforms with respect to the variables (t_0, z_0) and (X, Y, Z), respectively.

Throughout this paper constants are invariably denoted by C.

The space we choose for the profiles must contain oscillations with a discrete spectrum such as $U(\tau, T, X, Y, Z)e^{i\theta}$ and oscillations with a purely continuous spectrum. The spaces used here are a generalization to diffractive scales of those which have been introduced in [12] to describe Raman scattering.

DEFINITION 1.1. (i) *We denote by A_0^s the set of the functions defined on $\mathbb{R}^3_{X,Y,Z} \times \mathbb{R}^2_{t_0,z_0}$ with values in $\mathbb{C}^n$ whose Fourier transform with respect to (t_0, z_0) belongs to the set $\mathcal{BV}(\mathbb{R}^2_\xi, H^s(\mathbb{R}^3_{X,Y,Z})^n)$ of bounded variation Borel measures defined on $\mathbb{R}^2_\xi$ and with values in $H^s(\mathbb{R}^3)^n$. This space is endowed with the norm*

$$||\mathcal{V}||_{A_0^s} := |\mathcal{F}\mathcal{V}|_{\mathcal{BV}} \qquad \forall \mathcal{V} \in A_0^s;$$

(ii) *We denote by $E^s_{\tau *}$ the set of the functions defined on $[0, \tau^*] \times \mathbb{R}_T \times \mathbb{R}^3_{X,Y,Z} \times \mathbb{R}^2_{t_0,z_0}$ with values in $\mathbb{C}^n$ whose Fourier transform with respect to (t_0, z_0) belongs to $\mathcal{C}([0, \tau^*] \times \mathbb{R}_T, \mathcal{BV}(\mathbb{R}^2_\xi, H^s(\mathbb{R}^3_{X,Y,Z})^n))$. Moreover, for all $T \in \mathbb{R}$, we define*

$$||\mathcal{V}(T)||_{E^s_{\tau^*}} := \sup_{0 \le \tau \le \tau^*} |\mathcal{F}\mathcal{V}(\tau, T)|_{\mathcal{BV}} \qquad \forall \mathcal{V} \in E^s_{\tau^*};$$

(iii) *We denote by $A^s_{\tau^*}$ the subspace of $E^s_{\tau *}$ composed by all the functions of $E^s_{\tau *}$ bounded on $[0, \tau^*] \times \mathbb{R}_T$ and endow this space with the norm*

$$||\mathcal{V}||_{A^s_{\tau^*}} := \sup_{0 \le \tau \le \tau^*} \sup_{T \in \mathbb{R}} |\mathcal{F}\mathcal{V}(\tau, T)|_{\mathcal{BV}} = \sup_{T \in \mathbb{R}} ||\mathcal{V}(T)||_{E^s_{\tau^*}} \qquad \forall \mathcal{V} \in A^s_{\tau^*};$$

(iv) *We denote by $B^s_{\tau^*}$ the subspace of $A^s_{\tau^*}$ composed by all the functions of $A^s_{\tau^*}$ which do not depend on T.*

The well-known notion of the oscillatory spectrum of an (almost-)periodic function [9] can then be generalized as follows.

DEFINITION 1.2. *If $\mathcal{V} \in E^s_{\tau^*}$, then for all $(\tau, T) \in [0, \tau^*] \times \mathbb{R}$, the spectrum* Sp $\mathcal{V}(\tau, T)$ *of $\mathcal{V}(\tau, T)$ is the support of the Fourier transform $\mathcal{F}\mathcal{V}(\tau, T)$.*

We also define the spectrum of $\mathcal{V}$ as Sp $\mathcal{V} = \bigcup_{(\tau,T) \in [0,\tau^*] \times \mathbb{R}}$ Sp $\mathcal{V}(\tau, T)$.

The following proposition [12] states the main properties of these functional spaces.

PROPOSITION 1.3. (i) *The two normed spaces $(A_0^s, ||.||_{A_0^s})$ and $(A^s_{\tau^*}, ||.||_{A^s_{\tau^*}})$ are complete.*

(ii) *Any J-linear mapping G defined on $(\mathbb{C}^n)^J$ and with values in $\mathbb{C}^n$ extends to a continuous J-linear mapping defined on A_0^s (resp., $A^s_{\tau^*}$) and with values in A_0^s (resp., $A^s_{\tau^*}$). Moreover, there exists a constant $l > 0$ such that for all J-uplet $(\mathcal{V}_1, \dots, \mathcal{V}_J) \in A_0^s$ (resp., $A^s_{\tau^*}$), one has*

$$||G(\mathcal{V}_1, \dots, \mathcal{V}_J)|| \le l||\mathcal{V}_1|| \dots ||\mathcal{V}_J||,$$

where $||.||$ represents the norm of A_0^s (resp., $A^s_{\tau^*}$).

(iii) *Let $\mathcal{V}$ be in A^s_0 (resp., $A^s_{\tau^*}$). Then $\mathcal{V}$ is also in $\mathcal{C}(\mathbb{R}^5)^n$ (resp., $\mathcal{C}([0,\tau^*]\times\mathbb{R}^6)^n$). Moreover, $\mathcal{V}$ is bounded, and there exists a positive number l' such that*

$$||\mathcal{V}||_\infty \leq l'||\mathcal{V}||_{A^s_0} \quad (resp., \quad ||\mathcal{V}||_\infty \leq l'||\mathcal{V}||_{A^s_{\tau^*}}).$$

(iv) *Let $\mathcal{V} \in A^s_0$ (resp., $A^s_{\tau^*}$). Then the function v^ε defined on $\mathbb{R}^3$ (resp., $[0,\frac{\tau^*}{\varepsilon}]\times\mathbb{R}^3$) as $v^\varepsilon(X,Y,Z) = \mathcal{V}(X,Y,Z,0,\frac{Z}{\varepsilon})$ (resp., $v^\varepsilon(T,X,Y,Z) = \mathcal{V}(\varepsilon T,T,X,Y,Z,\frac{T}{\varepsilon},\frac{Z}{\varepsilon})$) belongs to $L^2(\mathbb{R}^3)^n$ (resp., $\mathcal{C}([0,\frac{\tau^*}{\varepsilon}],L^2(\mathbb{R}^3)^n)$). Moreover, one has*

$$||v^\varepsilon||_{L^2(\mathbb{R}^3)} \leq ||\mathcal{V}||_{A^s_0} \quad (resp., \sup_{0\leq T\leq\tau^*/\varepsilon} ||v^\varepsilon(T,.)||_{L^2(\mathbb{R}^3)} \leq ||\mathcal{V}||_{A^s_{\tau^*}}).$$

Examples.

Example 1. Oscillations with a discrete spectrum such as $U(\tau,T,X,Y,Z)e^{i\theta}$, with $\theta = \omega_l t_0 - k_l z_0$ and $U \in \mathcal{C}([0,\tau^*]\times\mathbb{R}_T, H^s(\mathbb{R}^3)^n)$, are in $E^s_{\tau^*}$. Indeed, taking the Fourier transform of such oscillations yields

$$\mathcal{F}(Ue^{i\theta}) = U\delta_{(\omega_l,-k_l)},$$

which belongs to $\mathcal{C}([0,\tau^*]\times\mathbb{R}_T, \mathcal{BV}(\mathbb{R}^2_\xi, H^s(\mathbb{R}^3_{X,Y,Z})^n))$.

Example 2. Let $\mathcal{M}$ be a submanifold of $\mathbb{R}^2$ and α an L^1 function defined on $\mathcal{M}$ and with values in $\mathcal{C}([0,\tau^*]\times\mathbb{R}_T, H^s(\mathbb{R}^3)^n)$. Then the density function [12] defined as

$$\mathcal{V}(\tau,T,X,Y,Z,t_0,z_0) = \int_{\mathcal{M}} e^{i(t_0,z_0)\cdot(\omega,k)}\alpha(\omega,k)(\tau,T,X,Y,Z)\sigma(d\omega,dk),$$

where σ denotes the Lebesgue measure of $\mathcal{M}$, is in $E^s_{\tau^*}$, and its oscillatory spectrum is $\mathcal{M}$.

1.3. Solving the Cauchy problem (1.1). We recall that (1.1) is written as

$$\begin{cases} L^\varepsilon(\partial)\mathbf{u}^\varepsilon + f(\mathbf{u}^\varepsilon) = 0, \\ \mathbf{u}^\varepsilon|_{T=0}(X,Y,Z) = \mathbf{u}^0_\varepsilon(X,Y,Z). \end{cases}$$

We now make precise the assumptions we make on $L^\varepsilon(\partial)$.

ASSUMPTION 1.1. *The system* (1.1) *is symmetric hyperbolic. More accurately, the operator $L^\varepsilon(\partial)$ can be written*

$$L^\varepsilon(\partial) = A_0\partial_T + A_1\partial_X + A_2\partial_Y + A_3\partial_Z + \frac{L_0}{\varepsilon},$$

where the A_i are real symmetric matrices and A_0 is strictly positive. Moreover the system (1.1) *is conservative in the sense that $(L_0)^* = -L_0$.*

Remark 1.2. Since A_0 is strictly positive, we can take $A_0^{-1/2}\mathbf{u}^\varepsilon$ as a new unknown. Multiplying (1.1) by $A_0(0)^{-1/2}$, the resulting system has the same properties as system (1.1) and satisfies $A_0(0) = Id$. *Thus herein, we always consider that $A_0 = Id$.*

The following hypothesis gives the kind of nonlinearity f we study here.

ASSUMPTION 1.2. *There exists a trilinear mapping F such that for all $u \in \mathbb{C}^n$, $f(u) = F(u,u,u)$.*

Remark 1.3. In this paper, we consider nonlinearities of order 3 since the two examples we gave in the last section belong to this class. This limitation on the order of the nonlinearity is only due to technical reasons, and the interested reader could easily generalize our results to nonlinearities of different orders.

The initial conditions for (1.1) must be general enough to allow a model of both large spectrums and ultrashort pulses. The spaces introduced in the previous part are adapted to such a general point of view, and we therefore consider initial conditions of the form

$$\mathbf{u}^0_\varepsilon(X,Y,Z) = \varepsilon^p \mathbf{U}^0\left(X,Y,Z,0,\frac{Z}{\varepsilon}\right), \tag{1.4}$$

with $\mathbf{U}^0 \in A^s_0$.

Choice of the size of the solutions. The choice of p is given [8] by the order of the nonlinearity, $p = 1/2$. With this choice, nonlinear and diffractive effects occur simultaneously.

The following theorem proves that the unique solution L^2 of the Cauchy problem (1.1) with initial condition (1.4) can be written using profiles from $B^s_{\tau^*}$.

THEOREM 1.4. *Let $R > 0$ and $\mathbf{U}^0$ in A^s_0 be such that $\|\mathbf{U}^0\|_{A^s_0} \le R$. There exists a positive real number $\tau^*_1 > 0$, which depends on $\mathbb{R}$ but not on ε, such that for all $\varepsilon > 0$, the Cauchy problem*

$$\begin{cases} L^\varepsilon(\partial)\mathbf{u}^\varepsilon + f(\mathbf{u}^\varepsilon) = 0, \\ \mathbf{u}^\varepsilon|_{T=0}(X,Y,Z) = \varepsilon^{\frac{1}{2}}\mathbf{U}^0(X,Y,Z,0,Z/\varepsilon) \end{cases}$$

has a unique solution $\mathbf{u}^\varepsilon$ in $\mathcal{C}([0,\frac{\tau^*_1}{\varepsilon}]\times\mathbb{R}^3)^n \cap \mathcal{C}([0,\frac{\tau^*_1}{\varepsilon}], L^2(\mathbb{R}^3)^n)$.

Moreover, $\mathbf{u}^\varepsilon$ can be written $\mathbf{u}^\varepsilon(T,X,Y,Z) := \varepsilon^{\frac{1}{2}}\mathbf{U}^\varepsilon(\varepsilon T, X, Y, Z, \frac{T}{\varepsilon}, \frac{Z}{\varepsilon})$, where $\mathbf{U}^\varepsilon \in B^s_{\tau^*_1}$ is uniquely determined by the so-called singular equation*,*

$$\begin{cases} \partial_\tau \mathbf{U}^\varepsilon + \varepsilon^{-1}(A_1\partial_X + A_2\partial_Y + A_3\partial_Z)\mathbf{U}^\varepsilon + \varepsilon^{-2}(\partial_{t_0} + A_3\partial_{z_0} + L_0)\mathbf{U}^\varepsilon + f(\mathbf{U}^\varepsilon) = 0, \\ \mathbf{U}^\varepsilon_{|\tau=0} = \mathbf{U}^0, \end{cases} \tag{1.5}$$

and for all $\varepsilon \in (0,1)$, $\mathbf{U}^\varepsilon$ satisfies the uniform bound $\|\mathbf{U}^\varepsilon\|_{B^s_{\tau^*_1}} \le 2R$.

Proof. The proof of this theorem is similar to the proof of the existence theorem of [12], and we give only a sketch of it. First, we prove that the existence of $\mathbf{u}^\varepsilon$ is a consequence of the existence of a profile $\mathbf{U}^\varepsilon$ satisfying (1.5). This latter result is obtained by Picard iterates using the following lemma, which gives linear estimates.

LEMMA 1.5. *Let $\mathcal{V}^0 \in A^s_0$ and $\mathcal{W} \in B^s_{\tau^*_1}$. The linear problem*

$$\begin{cases} \partial_\tau \mathcal{V} + \varepsilon^{-1}(A_1\partial_X + A_2\partial_Y + A_3\partial_Z)\mathcal{V} + \varepsilon^{-2}(\partial_{t_0} + A_3\partial_{z_0} + L_0)\mathcal{V} = \mathcal{W}, \\ \mathcal{V}_{|\tau=0} = \mathcal{V}^0 \end{cases}$$

has a unique solution in $B^s_{\tau^*_1}$. Moreover, one has

$$||\mathcal{V}||_{B^s_{\tau^*_1}} = ||\mathcal{V}^0||_{A^s_0} + \tau^*_1||\mathcal{W}||_{B^s_{\tau^*_1}}.$$

The existence of $\mathbf{u}^\varepsilon$ being established, one proves uniqueness using a classical L^2-uniqueness argument. ∎

1.4. General method. We seek an approximate solution u^ε of the exact solution $\mathbf{u}^\varepsilon$ of (1.1) using the tools of diffractive optics. The approximate solution u^ε is sought in the form

$$u^\varepsilon = \varepsilon^{\frac{1}{2}}\mathcal{U}^\varepsilon\left(\varepsilon T, T, X, Y, Z, \frac{T}{\varepsilon}, \frac{Z}{\varepsilon}\right), \qquad \text{with} \quad \mathcal{U}^\varepsilon = \mathcal{U}_0 + \varepsilon\mathcal{U}_1 + \varepsilon^2\mathcal{U}_2, \tag{1.6}$$

and $\mathcal{U}_0, \mathcal{U}_1, \mathcal{U}_2 \in E^s_{\tau^*}$.

The expansion of $L^\varepsilon(\partial)u^\varepsilon + f(u^\varepsilon)$ in powers of ε yields

$$L^\varepsilon(\partial)u^\varepsilon + f(u^\varepsilon) = \sum_{j=-1}^{7} \varepsilon^{\frac{1}{2}+j}\mathcal{R}_j(\tau, T, X, Y, Z, t_0, z_0)|_{\tau=\varepsilon T, t_0=T/\varepsilon, z_0=Z/\varepsilon}, \tag{1.7}$$

where

$$\begin{aligned}
\mathcal{R}_{-1}(\tau, T, X, Y, Z, t_0, z_0) &= i\mathcal{L}(D_{t_0,z_0})\mathcal{U}_0,\\
\mathcal{R}_0(\tau, T, X, Y, Z, t_0, z_0) &= i\mathcal{L}(D_{t_0,z_0})\mathcal{U}_1 + L_1(\partial)\mathcal{U}_0,\\
\mathcal{R}_1(\tau, T, X, Y, Z, t_0, z_0) &= i\mathcal{L}(D_{t_0,z_0})\mathcal{U}_2 + L_1(\partial)\mathcal{U}_1 + \partial_\tau\mathcal{U}_0 + f(\mathcal{U}_0),\\
\mathcal{R}_2(\tau, T, X, Y, Z, t_0, z_0) &= L_1(\partial)\mathcal{U}_2 + \partial_\tau\mathcal{U}_1 + \langle f(\mathcal{U}^\varepsilon)\rangle_{1/2+2},\\
\mathcal{R}_3(\tau, T, X, Y, Z, t_0, z_0) &= \partial_\tau\mathcal{U}_2 + \langle f(\mathcal{U}^\varepsilon)\rangle_{1/2+3},\\
\mathcal{R}_{j\geq 4}(\tau, T, X, Y, Z, t_0, z_0) &= \langle f(\mathcal{U}^\varepsilon)\rangle_{1/2+j},
\end{aligned} \tag{1.8}$$

with the notation

$$\begin{aligned}
&\mathcal{L}(D_{t_0,z_0}) := D_{t_0} + A_3 D_{z_0} + L_0/i, \qquad D_{t_0} = -i\partial_{t_0}, \quad D_{z_0} = -i\partial_{z_0}\\
&L_1(\partial) := \partial_T + A_1\partial_X + A_2\partial_Y + A_3\partial_Z := \partial_T + A(\partial_{X,Y,Z}),
\end{aligned}$$

while $\langle f(\mathcal{U}^\varepsilon)\rangle_k$ denotes the coefficient of the monomial ε^k in the expansion into powers of ε of $f(\mathcal{U}^\varepsilon)$.

Notation. We used the pseudodifferential notation $D_{t_0} = -i\partial_{t_0}$ and $D_{z_0} = -i\partial_{z_0}$ to define the operator $\mathcal{L}(D_{t_0,z_0})$. This explains the factor i which appears in front of it in expansion (1.8). Recalling that (ω, k) denote the dual variables of (t_0, z_0), the symbol of this operator reads $\mathcal{L}(\omega, k) = \omega Id + kA_3 + L_0/i$.

The strategy of diffractive optics consists of seeking $\mathcal{U}_0$, $\mathcal{U}_1$, and $\mathcal{U}_2$ in order to cancel the profiles $\mathcal{R}_m(\tau, T, X, Y, Z, t_0, z_0)$, $m = -1, 0, 1$. We then prove that the associated function u^ε given by (1.6) is indeed an approximate solution of (1.1) and give a stability theorem.

1.5. A few tools. The following definition introduces some concepts of diffractive optics.

DEFINITION 1.6. (i) *The characteristic variety associated to the operator $\mathcal{L}$ is the set*

$$\mathcal{C}_\mathcal{L} = \{(\omega, k) \in \mathbb{R}^2, \det(\mathcal{L}(\omega, k)) = \det(\omega Id + kA_3 + L_0/i) = 0\}.$$

(ii) *We denote by $\pi(\omega, k)$ the orthogonal projector onto* $\ker \mathcal{L}(\omega, k)$ *and by $\mathcal{L}^{-1}(\omega, k)$ the partial inverse of $\mathcal{L}(\omega, k)$ defined as*

$$\mathcal{L}^{-1}(\omega, k)\pi(\omega, k) = 0 \quad \text{and} \quad \mathcal{L}^{-1}(\omega, k)\mathcal{L}(\omega, k) = Id - \pi(\omega, k).$$

(iii) *Near every smooth point $(\underline{\omega}, \underline{k})$ of $\mathcal{C}_\mathcal{L}$, we denote by $\omega(k)$ a local parameterization of $\mathcal{C}_\mathcal{L}$.*

The following lemma expresses the resolubility condition of a linear equation with the tools introduced in the previous definition.

LEMMA 1.7. *Let $a, b \in \mathbb{C}^n$. Then the following two assertions are equivalent:*
(i) $\mathcal{L}(\omega, k)a = b$;
(ii) $\pi(\omega, k)b = 0$ *and* $(Id - \pi(\omega, k))a = \mathcal{L}^{-1}(\omega, k)b$.

We want to generalize these resolubility conditions to equations of type $\mathcal{L}(D_{t_0,z_0})\mathcal{V} = \mathcal{W}$, where $\mathcal{V}$ and $\mathcal{W}$ are in $E^s_{\tau *}$. Following [12], we first have to introduce the notion of $\mathcal{L}^{-1}$-regularity.

DEFINITION 1.8. *Let $\mathcal{V} \in E^s_{\tau^*}$ and $\mu(\tau, T) := \mathcal{F}\mathcal{V}(\tau, T)$. We say that $\mathcal{V}$ is $\mathcal{L}^{-1}$-regular if for all $(\tau, T) \in [0, \tau^*] \times \mathbb{R}_T$, $\mathcal{L}^{-1}$ is $\mu(\tau, T)$-integrable.*

We can now generalize Lemma 1.7 in the following way.

LEMMA 1.9. *Let $\mathcal{V}$ and $\mathcal{W}$ be in $E^s_{\tau^*}$. The following assertions are equivalent:*

(i) $\mathcal{L}(D_{t_0,z_0})\mathcal{V} = \mathcal{W}$;

(ii) $\pi(D_{t_0,z_0})\mathcal{W} = 0$, $\mathcal{W}$ *is* $\mathcal{L}^{-1}$*-regular, and* $(Id - \pi(D_{t_0,z_0}))\mathcal{V} = \mathcal{L}^{-1}(D_{t_0,z_0})\mathcal{W}$.

Remark 1.4. The $\mathcal{L}^{-1}$-regularity condition may not be satisfied in the physical phenomena we are interested in here. Indeed, the mapping $(\omega, k) \to \mathcal{L}^{-1}(\omega, k)$ is not bounded at the neighborhood of the origin. When low frequencies are excluded, as in most applications in optics, $\mathcal{L}^{-1}$ remains bounded for the frequencies considered, and the $\mathcal{L}^{-1}$-regularity condition is easily satisfied. But when low frequencies are allowed, as they have to be here, $\mathcal{L}^{-1}$ effectively blows up and the $\mathcal{L}^{-1}$-regularity condition is in general not satisfied. In that case, we have to use tools similar to those introduced by Alterman [1].

It is also interesting to decompose the profiles of $E^s_{\tau^*}$ into a discrete spectrum (sinusoidal) component and a purely continuous one. Such a decomposition is assured by the foregoing proposition.

PROPOSITION 1.10. *Let $\mathcal{V} \in A^s_0$ (resp., $E^s_{\tau^*}$). The profile $\mathcal{V}$ is written uniquely as $\mathcal{V} = \mathcal{V}_I + \mathcal{V}_{II}$, with $\mathcal{V}_I$, $\mathcal{V}_{II} \in A^s_0$ (resp., $E^s_{\tau^*}$) such that*

(i) $\mathcal{V}_I$ *(resp., $\mathcal{V}_I(\tau, T, .)$ for all $(\tau, T) \in [0, \tau^*] \times \mathbb{R}_T$) has a discrete spectrum;*

(ii) $\mathcal{V}_{II}$ *has a purely continuous spectrum, i.e., every point of $\mathbb{R}^2$ has zero measure for $\mathcal{F}\mathcal{V}_{II}$ (resp., for $\mathcal{F}\mathcal{V}_{II}(\tau, T, .)$ for all $(\tau, T) \in [0, \tau^*] \times \mathbb{R}_T$).*

Notation. From now on, and for every profile $\mathcal{V}$ of A^s_0 or $E^s_{\tau^*}$, we denote by $\mathcal{V}_I$ the component with a discrete spectrum and by $\mathcal{V}_{II}$ the component with a purely continuous one.

Proof. First, the existence of the decomposition of a profile $\mathcal{V} \in E^s_{\tau^*}$ is proved. We introduce

$$\mu(\tau, T) := \mathcal{F}\mathcal{V}(\tau, T) \qquad \forall (\tau, T) \in [0, \tau^*] \times \mathbb{R}_T,$$

and $S_{\tau,T}$ the set of points with nonzero measure for $\mu(\tau, T)$,

$$S_{\tau,T} = \{p \in \mathbb{R}^2, \mu(\tau, T)(\{p\}) \neq 0\}.$$

We decompose $\mu(\tau, T)$ in the form

$$\mu(\tau, T) = \mathbb{I}_{S_{\tau,T}}\mu(\tau, T) + (1 - \mathbb{I}_{S_{\tau,T}})\mu(\tau, T),$$

and the proposition will be proved if we can show that $\mathcal{F}^{-1}(\mathbb{I}_{S_{\tau,T}}\mu(\tau, T))$ and $\mathcal{F}^{-1}[(1 - \mathbb{I}_{S_{\tau,T}})\mu(\tau, T)]$ are in $E^s_{\tau^*}$. Indeed, one would then have for all (τ, T),

$$\mathcal{V}_I(\tau, T) = \mathcal{F}^{-1}(\mathbb{I}_{S_{\tau,T}}\mu(\tau, T)) \quad \text{and} \quad \mathcal{V}_{II}(\tau, T) = \mathcal{F}^{-1}([1 - \mathbb{I}_{S_{\tau,T}}]\mu(\tau, T)).$$

It is clear that for all $(\tau, T) \in [0, \tau^*] \times \mathbb{R}_T$, the measures $\mathbb{I}_{S_{\tau,T}}\mu(\tau, T)$ and $(1 - \mathbb{I}_{S_{\tau,T}})\mu(\tau, T)$ belong to $\mathcal{BV}(\mathbb{R}^2_\xi, H^s(\mathbb{R}^3)^n)$. The main difficulty is to prove the continuous dependence on (τ, T) of these two measures. As the mapping $(\tau, T) \mapsto \mu(\tau, T)$ is continuous by assumption, it is sufficient to prove that the mapping

$$\begin{aligned} [0, \tau^*] \times \mathbb{R}_T &\longrightarrow \mathcal{BV}(\mathbb{R}^2, H^s(\mathbb{R}^3)^n), \\ (\tau, T) &\longmapsto \mathbb{I}_{S_{\tau,T}}\mu(\tau, T) \end{aligned}$$

is continuous, i.e., that $\mathbb{I}_{S_{\tau,T}}\mu(\tau,T) - \mathbb{I}_{S_{\tau',T'}}\mu(\tau',T')$ tends to 0 in $\mathcal{BV}(\mathbb{R}^2_\xi, H^s(\mathbb{R}^3)^n)$ when (τ',T') tends to (τ,T). One has

$$\mathbb{I}_{S_{\tau,T}}\mu(\tau,T) - \mathbb{I}_{S_{\tau',T'}}\mu(\tau',T') = (\mathbb{I}_{S_{\tau,T}} - \mathbb{I}_{S_{\tau',T'}})\mu(\tau,T) + \mathbb{I}_{S_{\tau',T'}}(\mu(\tau,T) - \mu(\tau',T')).$$

The second term of the right-hand side of this equation tends to 0 when (τ',T') tends to (τ,T) thanks to the continuity of the mapping $(\tau,T) \mapsto \mu(\tau,T)$. Moreover, the first term of the right-hand side of the equation reads

$$(\mathbb{I}_{S_{\tau,T}} - \mathbb{I}_{S_{\tau',T'}})\mu(\tau,T) = \mathbb{I}_{S_{\tau,T}\setminus S_{\tau',T'}}\mu(\tau,T) - \mathbb{I}_{S_{\tau',T'}\setminus S_{\tau,T}}\mu(\tau,T).$$

But one has $\mathbb{I}_{S_{\tau',T'}\setminus S_{\tau,T}}\mu(\tau,T) = 0$, because if $p \in S_{\tau',T'} \setminus S_{\tau,T}$, then $\mu(\tau,T)(\{p\}) = 0$. On the other hand, one has $\mathbb{I}_{S_{\tau,T}\setminus S_{\tau',T'}}(\{p\}) \to 0$ for all $p \in \mathbb{R}^2$ when $(\tau',T') \to (\tau,T)$. Indeed, for all $p \in S_{\tau,T}$, $\mu(\tau,T)(\{p\}) \neq 0$. By continuity of the mapping $(\tau,T) \mapsto \mu(\tau,T)$, one has $\mu(\tau',T')(\{p\}) \longrightarrow \mu(\tau,T)(\{p\})$ in $H^s(\mathbb{R}^3)^n$ when $(\tau',T') \to (\tau,T)$. Consequently, $\mu(\tau',T')(\{p\}) \neq 0$ if (τ',T') is close enough to (τ,T). In other words, $p \in S_{\tau',T'}$, and hence $p \notin S_{\tau,T} \setminus S_{\tau',T'}$. The proof of the continuous time dependence is then achieved by a dominated convergence argument.

The existence of the decomposition for every profile of $E^s_{\tau *}$, and a fortiori of A^s_0, is thus proved. The proof of the uniqueness of the decomposition is straightforward. □

1.6. Organization of the paper. We first address in section 2 general dispersive hyperbolic systems. Section 2.1 is devoted to the derivation of the profile equations under a simplifying assumption of absence of low frequencies. In section 2.2, we perform a sharp analysis of the nonlinearities found for the profile equations. We show that many of these nonlinearities vanish, which is of crucial importance for the resolubility theorems given in section 2.3. The fact that the approximate solution associated to the profiles found in the previous sections converges towards the exact solution of (1.1) is then proved in section 2.4. The general case (presence of low frequencies) is addressed in section 2.5. Using Alterman's technique of infrared cutoffs, we use the results of the previous sections to give profile equations in the general case, as well as a stability theorem which generalizes the one given in section 2.4.

In section 3, we treat the nondispersive case. Since the methods used in this section are the same as in the dispersive case, we do not extend the proofs. However, the main difference between both cases is by itself interesting enough to justify this section: *nonlinear interactions between components with a purely continuous spectrum can be observed in the nondispersive case*, while they are negligible in the dispersive case.

The two physical examples used as guidelines throughout this paper are studied in section 4. These examples are lasers with large spectrums and short pulses and illustrate the notable simplifications yielded by our general theory with respect to direct computations.

Finally, an intermediate case between dispersive and nondispersive systems, called weakly dispersive, is briefly commented on in section 5. In particular, we find that for large-spectrum lasers, equations for the continuous spectrum component are still linear but, as opposed to the dispersive case, coupled with the discrete spectrum component.

2. Dispersive case. In this part, we consider problems of type (1.1) which are dispersive. More precisely, we suppose that the following assumption is satisfied.

ASSUMPTION 2.1. *One has* $\{0, \pm(\omega_l, -k_l)\} \subset \mathcal{C}_\mathcal{L}$*, but for all* $j \in Z \setminus \{0, \pm 1\}$*, the point* $j(\omega_l, -k_l)$ *is not on* $\mathcal{C}_\mathcal{L}$*. We also assume that* $\mathcal{C}_\mathcal{L}$ *is a union of smooth curves which are never parallel, asymptotic, nor tangent to one another, and which intersect only on the vertical axis* $(O\omega)$.

Remark 2.1. The different nonlinear models whose linearization gives the Maxwell–Lorentz equations (see the last section for an example) do not exactly satisfy this assumption since $\mathcal{C}_\mathcal{L}$ then contains three horizontal lines, which are a fortiori parallel. Moreover, two of these lines are also tangent to curved sheets of the characteristic variety. However, this is not important since the horizontal lines are excluded by the divergence-free conditions one has to add to these systems. Therefore, the Maxwell–Lorentz model, and its nonlinear versions, fall into the range of this assumption.

2.1. The ansatz in the absence of infrared frequencies. As we have said in Remark 1.4, low frequencies make the analysis far more difficult because $\mathcal{L}^{-1}$-regularity of the profiles may fail. That is why in this section we focus on profiles $\mathcal{U}_0$ whose spectrum is outside the band $\{(\omega, k), |k| \leq \delta\}$. More precisely, we assume throughout this section that the continuous spectrum component of the leading term $\mathcal{U}_0$ satisfies the following assumption.

ASSUMPTION 2.2. *The spectrum* $\mathrm{Sp}\ \mathcal{U}_{0,II}$ *of the continuous spectrum component of* $\mathcal{U}_0$ *is in* $\{(\omega, k), |k| > \delta\}$*, where* $\delta > 0$.

The following lemma makes the link between absence of low frequencies and $\mathcal{L}^{-1}$-regularity.

LEMMA 2.1. *Let* $\mathcal{V}_{II}$ *be a profile of* $E^s_{\tau^*}$ *such that* $\mathrm{Sp}\ \ \mathcal{V}_{II} \subset \mathcal{C}_\mathcal{L}$.

If, moreover, $\mathrm{Sp}\ \mathcal{V}_{II} \subset \{(\omega, k), |k| > \delta\}$*, then* $\mathcal{V}_{II}$ *is* $\mathcal{L}^{-1}$*-regular, and for all* $T \in \mathbb{R}$*,*

$$\|\mathcal{L}^{-1}\mathcal{V}_{II}(T)\|_{E^s_{\tau^*}} \leq \frac{C}{\delta}\|\mathcal{V}_{II}(T)\|_{E^s_{\tau^*}}.$$

In particular, if $\mathcal{V}_{II} \in A^s_{\tau^*}$*, one has*

$$\|\mathcal{L}^{-1}\mathcal{V}_{II}\|_{A^s_{\tau^*}} \leq \frac{C}{\delta}\|\mathcal{V}_{II}\|_{A^s_{\tau^*}}.$$

Proof. For all $(\omega, k) \in \mathbb{R}^2$, $\mathcal{L}^{-1}(\omega, k)$ reads

$$\mathcal{L}^{-1}(\omega, k) = \sum_{j, \omega_j(k) \neq \omega} \frac{1}{\omega - \omega_j(k)} \pi(\omega_j(k), k),$$

where the ω_j are parameterizations of the different sheets of $\mathcal{C}_\mathcal{L}$. If $(\omega, k) \in \mathcal{C}_\mathcal{L}$, i.e., if there exists j_0 such that $(\omega, k) = (\omega_{j_0}(k), k)$, then

$$\mathcal{L}^{-1}(\omega, k) = \mathcal{L}^{-1}(\omega_{j_0}(k), k) = \sum_{j \neq j_0} \frac{1}{\omega_{j_0}(k) - \omega_j(k)} \pi(\omega_j(k), k). \tag{2.1}$$

Saying that $\mathcal{V}_{II}$ is $\mathcal{L}^{-1}$-regular means that for all $(\tau, T) \in [0, \tau^*] \times \mathbb{R}$, $\mathcal{L}^{-1}(\omega, k)$ is integrable for $\mu(\tau, T) := \mathcal{F}\mathcal{V}_{II}(\tau, T)$. As we have supposed that $\mathrm{Sp}\ \mathcal{V}_{II} \subset \mathcal{C}_\mathcal{L}$, we must prove that the expression given by (2.1) is integrable for $\mu(\tau, T)$.

Since we supposed in Assumption 2.1 that the different sheets of $\mathcal{C}_\mathcal{L}$ are not asymptotic, (2.1) is bounded for large $|k|$. The only points where this expression is not bounded are those where the different sheets intersect. Thanks to Assumption 2.1,

these points are all on the axis $(O\omega)$. We now consider what happens in the neighborhood of such a point. Consider j and j_0 such that $\lim_{k\to 0}\omega_{j_0}(k)=\lim_{k\to 0}\omega_j(k)=\omega_0$. Near 0, one has

$$\frac{1}{\omega_{j_0}(k)-\omega_j(k)}=\frac{1}{(\omega_{j_0}(k)-\omega_0)-(\omega_j(k)-\omega_0)}\sim\frac{1}{k(v_0-v)},$$

where $v_0=\lim_{k\to 0}\omega'_{j_0}(k)$ and $v=\lim_{k\to 0}\omega'_j(k)$. We know that $v_0-v\neq 0$ since Assumption 2.1 assures us that two different sheets are never tangent.

Therefore, the expression given by (2.1) can be bounded for $|k|>\delta$ by C/δ, which yields both the $\mathcal{L}^{-1}$-regularity result for $\mathcal{V}_{II}$ and the estimate of the lemma. ∎

2.1.1. Annihilating $\mathcal{R}_{-1}$. Annihilating the $\varepsilon^{-\frac{1}{2}}$ term in expansion (1.7) is equivalent to $\mathcal{L}(D_{t_0,z_0})\mathcal{U}_0=0$. Thanks to Lemma 1.7, this equation is equivalent to the *polarization condition*

$$\pi(D_{t_0,z_0})\mathcal{U}_0=\mathcal{U}_0.$$

Moreover, thanks to Proposition 1.10, $\mathcal{U}_0$ can be decomposed in the form $\mathcal{U}_0=\mathcal{U}_{0,I}+\mathcal{U}_{0,II}$, where $\mathcal{U}_{0,I}$ has a discrete spectrum and $\mathcal{U}_{0,II}$ a purely continuous one.

Looking for $\mathcal{U}_{0,I}$ of the form $\mathcal{U}_{0,I,1}(\tau,T,X,Y,Z)e^{i\theta}$ + c.c., the polarization condition $\pi(D_{t_0,z_0})\mathcal{U}_0=\mathcal{U}_0$ gives

$$\pi(\omega_l,-k_l)\mathcal{U}_{0,I,1}=\mathcal{U}_{0,I,1}\quad\text{and}\quad\pi(D_{t_0,z_0})\mathcal{U}_{0,II}=\mathcal{U}_{0,II}. \tag{2.2}$$

Remark 2.2. The notation *c.c.* used above denotes the complex conjugate of the preceding expression. As we are concerned with real-valued solutions of (1.1) and (1.5), we always assume that in the Fourier expansion of the discrete spectrum components, one has $\mathcal{U}_{j,I,k}=\overline{\mathcal{U}_{j,I,-k}}$.

2.1.2. Annihilating $\mathcal{R}_0$. Annihilating the $\varepsilon^{\frac{1}{2}}$ term in expansion (1.7) reads $i\mathcal{L}(D_{t_0,z_0})\mathcal{U}_1+L_1(\partial)\mathcal{U}_0=0$.

As for $\mathcal{U}_0$, decompose $\mathcal{U}_1$ in the form $\mathcal{U}_1=\mathcal{U}_{1,I}+\mathcal{U}_{1,II}$, where $\mathcal{U}_{1,I}$ has a discrete spectrum and $\mathcal{U}_{1,II}$ a purely continuous one. We also look for $\mathcal{U}_{1,I}$ in the form $\mathcal{U}_{1,I,1}(\tau,T,X,Y,Z)e^{i\theta}+c.c.$, so that the equation $\mathcal{R}_0=0$ may read

$$\begin{cases} i\mathcal{L}(\omega_l,-k_l)\mathcal{U}_{1,I,1}+L_1(\partial)\mathcal{U}_{0,I,1}=0,\\ i\mathcal{L}(D_{t_0,z_0})\mathcal{U}_{1,II}+L_1(\partial)\mathcal{U}_{0,II}=0.\end{cases} \tag{2.3}$$

With the polarization condition (2.2) and Lemma 1.7, the first equation of (2.3) is equivalent to

$$\begin{cases} \pi(\omega_l,-k_l)L_1(\partial)\pi(\omega_l,-k_l)\mathcal{U}_{0,I,1}=0,\\ (Id-\pi(\omega_l,-k_l))\mathcal{U}_{1,I,1}=i\mathcal{L}^{-1}(\omega_l,-k_l)A(\partial_{X,Y,Z})\mathcal{U}_{0,I,1}.\end{cases} \tag{2.4}$$

Since Assumption 2.2 and Lemma 2.1 assure us that $A(\partial_{X,Y,Z})\mathcal{U}_{0,II}$ is $\mathcal{L}^{-1}$-regular, we know, thanks to Lemma 1.9, that the second equation of (2.3) is equivalent to

$$\begin{cases} \pi(D_{t_0,z_0})L_1(\partial)\pi(D_{t_0,z_0})\mathcal{U}_{0,II}=0,\\ (Id-\pi(D_{t_0,z_0}))\mathcal{U}_{1,II}=i\mathcal{L}^{-1}(D_{t_0,z_0})A(\partial_{X,Y,Z})\mathcal{U}_{0,II}.\end{cases} \tag{2.5}$$

Remark 2.3. At this stage, only the component $(Id-\pi(\omega_l,-k_l))\mathcal{U}_{1,I,1}$ is determined. We can therefore choose to take the other component equal to zero, i.e.,

$$\pi(\omega_l,-k_l)\mathcal{U}_{1,I,1}=0. \tag{2.6}$$

We cannot do the same thing for $\pi(D_{t_0,z_0})\mathcal{U}_{1,II}$ because this component will play an important role in the solvability of the profile equations.

2.1.3. Annihilating $\mathcal{R}_1$. Annihilating the $\varepsilon^{\frac{3}{2}}$ term in expansion (1.7) yields $i\mathcal{L}(D_{t_0,z_0})\mathcal{U}_2+L_1(\partial)\mathcal{U}_1+\partial_\tau\mathcal{U}_0+f(\mathcal{U}_0)=0$. Thanks to Proposition 1.10, this equation can be decomposed into a discrete spectrum component,

$$-i\mathcal{L}(\omega_l,-k_l)\mathcal{U}_{2,I}=L_1(\partial)\mathcal{U}_{1,I}+\partial_\tau\mathcal{U}_{0,I}+f(\mathcal{U}_0)_I, \tag{2.7}$$

and a purely continuous one,

$$-i\mathcal{L}(D_{t_0,z_0})\mathcal{U}_{2,II}=L_1(\partial)\mathcal{U}_{1,II}+\partial_\tau\mathcal{U}_{0,II}+f(\mathcal{U}_0)_{II}. \tag{2.8}$$

The nonlinearity $f(\mathcal{U}_0)_I$ in (2.7) is given by $f(\mathcal{U}_0)_I=f(\mathcal{U}_{0,I})$, which is a trigonometric polynomial,

$$f(\mathcal{U}_{0,I})=f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}})e^{i\theta}+f(\mathcal{U}_{0,I,1})e^{i3\theta}+\text{c.c.} \tag{2.9}$$

Since the third harmonic is created by the nonlinearity, we must seek $\mathcal{U}_{2,I}$ in the form

$$\mathcal{U}_{2,I}(\tau,T,X,Y,Z,\theta)=\mathcal{U}_{2,I,1}(\tau,T,X,Y,Z)e^{i\theta}+\mathcal{U}_{2,I,3}(\tau,T,X,Y,Z)e^{i3\theta}+\text{c.c.}$$

According to Assumption 2.1, the component $\mathcal{U}_{2,I,3}$ can be found by elliptic inversion since $\mathcal{L}(3\omega_l,-3k_l)$ is then nonsingular,

$$\mathcal{U}_{2,I,3}=\mathcal{L}(3\omega_l,-3k_l)^{-1}f(\mathcal{U}_{0,I,1}). \tag{2.10}$$

The remaining component $\mathcal{U}_{2,I,1}$ satisfies

$$i\mathcal{L}(\omega_l,-k_l)\mathcal{U}_{2,I,1}+L_1(\partial)\mathcal{U}_{1,I,1}+\partial_\tau\mathcal{U}_{0,I,1}+f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}})=0,$$

and thanks to Lemma 1.7, we get

$$\begin{cases}\pi(\omega_l,-k_l)\big(L_1(\partial)\mathcal{U}_{1,I,1}+\partial_\tau\mathcal{U}_{0,I,1}+f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}})\big)=0,\\ (Id-\pi(\omega_l,-k_l))\mathcal{U}_{2,I,1}=i\mathcal{L}^{-1}(\omega_l,-k_l)\big(L_1(\partial)\mathcal{U}_{1,I,1}+\partial_\tau\mathcal{U}_{0,I,1}+f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}})\big).\end{cases}$$

Using the polarization condition (2.2) and equations (2.4) and (2.6), the first equation of the previous system reads

$$\begin{aligned}&\partial_\tau\mathcal{U}_{0,I,1}+i\pi(\omega_l,-k_l)A(\partial_{X,Y,Z})\mathcal{L}^{-1}(\omega_l,-k_l)A(\partial_{X,Y,Z})\pi(\omega_l,-k_l)\mathcal{U}_{0,I,1}\\ &\qquad+\pi(\omega_l,-k_l)f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}})=0,\end{aligned} \tag{2.11}$$

while the second equation gives $(Id-\pi(\omega_l,-k_l))\mathcal{U}_{2,I,1}$ in terms of $\mathcal{U}_{0,I,1}$,

$$\begin{aligned}&(I-\pi(\omega_l,-k_l))\mathcal{U}_{2,I,1}\\ &=i\mathcal{L}^{-1}(\omega_l,-k_l)\big(iL_1(\partial)\mathcal{L}^{-1}(\omega_l,-k_l)A(\partial_{X,Y,Z})\mathcal{U}_{0,I,1}+f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}})\big).\end{aligned} \tag{2.12}$$

While the analysis of the discrete and the continuous spectrum components was formally the same in section 2.1.2, this is no longer true here. This is due to the fact that Assumption 2.2 cannot ensure that the right-hand side of (2.8) is $\mathcal{L}^{-1}$-regular. Therefore, Lemma 1.9 cannot be invoked to solve this equation.

Because of this difficulty, we cannot in general find $\mathcal{U}_2$ in $E^s_{\tau^*}$ such that $\mathcal{R}_1=0$. However, $\mathcal{U}_2\in E^s_{\tau^*}$ may be chosen in such a way that $\mathcal{R}_1$ is very small. We first introduce Alterman's infrared cutoff filter.

Definition 2.2. *Let ψ be defined as follows: $\psi(k)=1$ for $|k|>1$ and $\psi(k)=0$ otherwise. For all $k\in\mathbb{R}$, $\psi^\delta(k)$ is defined as $\psi^\delta(k)=\psi(k/\delta)$, where $\delta>0$ is the same as in Assumption* 2.2.

Remark 2.4. The function ψ introduced above is not smooth, while Alterman's filter is smooth. Since our framework allows it, we have made this choice in order to lighten a few equations. In particular, we have the equivalence $\mathrm{Sp}\ \mathcal{V} \subset \{(\omega,k), |k| > \delta\} \iff \psi^\delta(D_{z_0})\mathcal{V} = \mathcal{V}$.

Since no condition has been found at this stage on $\pi(D_{t_0,z_0})\mathcal{U}_{1,II}$, we can impose a condition of absence of low frequencies on this component and, more precisely, that

$$\psi^\delta(D_{z_0})\pi(D_{t_0,z_0})\mathcal{U}_{1,II} = \pi(D_{t_0,z_0})\mathcal{U}_{1,II}. \tag{2.13}$$

Instead of solving (2.8), we solve the approximate equation $(2.8)_\delta$ defined as

$$-i\mathcal{L}(D_{t_0,z_0})\mathcal{U}_{2,II} = L_1(\partial)\mathcal{U}_{1,II} + \partial_\tau \mathcal{U}_{0,II} + \psi^\delta(D_{z_0})f(\mathcal{U}_0)_{II}.$$

Using (2.5), this equation reads

$$\begin{aligned} -i\mathcal{L}(D_{t_0,z_0})\mathcal{U}_{2,II} = {} & iL_1(\partial)\mathcal{L}^{-1}(D_{t_0,z_0})A(\partial_{X,Y,Z})\mathcal{U}_{0,II} + L_1(\partial)\pi(D_{t_0,z_0})\mathcal{U}_{1,II} \\ & + \partial_\tau \mathcal{U}_{0,II} + \psi^\delta(D_{z_0})f(\mathcal{U}_0)_{II}. \end{aligned} \tag{2.14}$$

Thanks to the presence of the filter ψ^δ, to (2.13), and to Assumption 2.2, the right-hand side of this equation is $\mathcal{L}^{-1}$-regular, so that (2.14) is equivalent to

$$\begin{aligned} & \partial_\tau \mathcal{U}_{0,II} + i\pi(D_{t_0,z_0})A(\partial_{X,Y,Z})\mathcal{L}^{-1}(D_{t_0,z_0})A(\partial_{X,Y,Z})\pi(D_{t_0,z_0})\mathcal{U}_{0,II} \\ & \quad + \pi(D_{t_0,z_0})L_1(\partial)\pi(D_{t_0,z_0})\mathcal{U}_{1,II} \\ & \quad + \pi(D_{t_0,z_0})\psi^\delta(D_{z_0})[f(\mathcal{U}_0)]_{II} = 0 \end{aligned} \tag{2.15}$$

and

$$\begin{aligned} (Id - \pi(D_{t_0,z_0}))\mathcal{U}_{2,II} = {} & -\mathcal{L}^{-1}(D_{t_0,z_0})L_1(\partial)\mathcal{L}^{-1}(D_{t_0,z_0})A(\partial_{X,Y,Z})\mathcal{U}_{0,II} \\ & + i\mathcal{L}^{-1}(D_{t_0,z_0})A(\partial_{X,Y,Z})\pi(D_{t_0,z_0})\mathcal{U}_{1,II} \\ & + i\mathcal{L}^{-1}(D_{t_0,z_0})\psi^\delta(D_{z_0})f(\mathcal{U}_0)_{II}. \end{aligned} \tag{2.16}$$

Remark 2.5. (i) As said above, $\mathcal{R}_1 = 0$ is not solved exactly. If we can find $\mathcal{U}_0$, $\mathcal{U}_1$, and $\mathcal{U}_2$ solutions of (2.2), (2.4)–(2.6), (2.10), (2.11)–(2.12), (2.13), and (2.15)–(2.16), we then have $\mathcal{R}_{-1} = \mathcal{R}_0 = 0$ but only

$$\mathcal{R}_1 = (1 - \psi^\delta(D_{z_0}))f(\mathcal{U}_0)_{II}. \tag{2.17}$$

We will see later that this term tends towards zero as $\delta \to 0$.

(ii) Only the components $(Id - \pi(\omega_l, -k_l))\mathcal{U}_{2,I,1}$ and $(Id - \pi(D_{t_0,z_0}))\mathcal{U}_{2,II}$ of $\mathcal{U}_{2,I,1}$ and $\mathcal{U}_{2,II}$ are determined. We can therefore choose to take

$$\pi(\omega_l, -k_l)\mathcal{U}_{2,I,1} = 0 \quad \text{and} \quad \pi(D_{t_0,z_0})\mathcal{U}_{2,II} = 0. \tag{2.18}$$

2.1.4. Simplification of the profile equations. According to the previous results, one has

$$\mathcal{U}_0(\tau,T,X,Y,Z,t_0,z_0) = \mathcal{U}_{0,I,1}(\tau,T,X,Y,Z)e^{i\theta} + c.c. + \mathcal{U}_{0,II}(\tau,T,X,Y,Z,t_0,z_0),$$

with

$$\pi(\omega_l, -k_l)\mathcal{U}_{0,I,1} = \mathcal{U}_{0,I,1} \quad \text{and} \quad \pi(D_{t_0,z_0})\mathcal{U}_{0,II} = \mathcal{U}_{0,II}. \tag{2.19}$$

Moreover $\mathcal{U}_{0,I,1}$ must satisfy

$$(2.20)\quad \begin{cases} \pi(\omega_l,-k_l)L_1(\partial)\pi(\omega_l,-k_l)\mathcal{U}_{0,I,1}=0,\\ \partial_\tau\mathcal{U}_{0,I,1}+i\pi(\omega_l,-k_l)A(\partial_{X,Y,Z})\mathcal{L}^{-1}(\omega_l,-k_l)A(\partial_{X,Y,Z})\pi(\omega_l,-k_l)\mathcal{U}_{0,I,1}\\ \qquad +\pi(\omega_l,-k_l)f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}})=0, \end{cases}$$

and the purely continuous spectrum component $\mathcal{U}_{0,II}$ must satisfy

$$(2.21)\quad \begin{cases} \pi(D_{t_0,z_0})L_1(\partial)\pi(D_{t_0,z_0})\mathcal{U}_{0,II}=0,\\ \partial_\tau\mathcal{U}_{0,II}+i\pi(D_{t_0,z_0})A(\partial_{X,Y,Z})\mathcal{L}^{-1}(D_{t_0,z_0})A(\partial_{X,Y,Z})\pi(D_{t_0,z_0})\mathcal{U}_{0,II}\\ \qquad +\pi(D_{t_0,z_0})L_1(\partial)\pi(D_{t_0,z_0})\mathcal{U}_{1,II}+\pi(D_{t_0,z_0})\psi^\delta(D_{z_0})[f(\mathcal{U}_0)]_{II}=0, \end{cases}$$

where we recall that $\pi(D_{t_0,z_0})\mathcal{U}_{1,II}$ must satisfy the condition of absence of low frequencies (2.13).

Before these systems are simplified, a new symbol, $\mathcal{M}$, is introduced.

DEFINITION 2.3. *We denote by $\mathcal{M}(\omega,K)$ the symbol defined for all $(\omega,K)\in\mathbb{R}^{1+3}$ as*

$$\mathcal{M}(\omega,K)=wId+K_1A_1+K_2A_2+K_3A_3+L_0/i,$$

where $K=(K_1,K_2,K_3)$.

We also define the orthogonal projector $\pi_{\mathcal{M}}(\omega,K)$ onto $\ker\mathcal{M}(\omega,K)$, the partial inverse $\mathcal{M}^{-1}(\omega,K)$ of $\mathcal{M}(\omega,k)$, and denote by $\mathcal{C}_{\mathcal{M}}$ the characteristic variety associated to $\mathcal{M}$, i.e., the set of points (ω,K) where $\mathcal{M}(\omega,K)$ is singular.

We also make the following assumption, satisfied, for instance, by Maxwell systems in isotropic media.

ASSUMPTION 2.3. *$\mathcal{C}_{\mathcal{M}}$ is axisymmetric around $(O\omega)$.*

We can now state the following proposition.

PROPOSITION 2.4. *Suppose that Assumption* 2.3 *is satisfied and that $\{(\omega_l,-k_l)\}$ is a smooth point of $\mathcal{C}_{\mathcal{L}}$. One then has the following:*

(i) $\pi(\omega_l,-k_l)L_1(\partial)\pi(\omega_l,-k_l)=\pi(\omega_l,-k_l)(\partial_T+\omega'(k_l)\partial_Z)$;

(ii)
$$\begin{aligned}&\pi(\omega_l,-k_l)A(\partial_{X,Y,Z})\mathcal{L}^{-1}(\omega_l,-k_l)A(\partial_{X,Y,Z})\pi(\omega_l,-k_l)\\ &\quad=\tfrac{\omega'(k_l)}{2k_l}\pi(\omega_l,-k_l)(\partial_X^2+\partial_Y^2)+\tfrac{\omega''(k_l)}{2}\pi(\omega_l,-k_l)\partial_Z^2.\end{aligned}$$

(iii) *If $\mathcal{V}_{II}$ is a profile with a purely continuous spectrum, then one also has*

$$\pi(D_{t_0,z_0})L_1(\partial)\pi(D_{t_0,z_0})\mathcal{V}_{II}=\pi(D_{t_0,z_0})\big(\partial_T-\omega'(D_{z_0})\partial_Z\big)\mathcal{V}_{II}.$$

(iv) *If, moreover, $A(\partial_{X,Y,Z})\pi(D_{t_0,z_0})\mathcal{V}_{II}$ is $\mathcal{L}^{-1}$-regular, then*

$$\begin{aligned}&\pi(D_{t_0,z_0})A(\partial_{X,Y,Z})\mathcal{L}^{-1}(D_{t_0,z_0})A(\partial_{X,Y,Z})\pi(D_{t_0,z_0})\mathcal{V}_{II}\\ &=\frac{\omega'(D_{z_0})}{2D_{z_0}}\pi(D_{t_0,z_0})(\partial_X^2+\partial_Y^2)\mathcal{V}_{II}+\frac{\omega''(D_{z_0})}{2}\partial_Z^2\mathcal{V}_{II}.\end{aligned}$$

Proof. If $(\underline{\omega},\underline{K})$ is a smooth point of $\mathcal{C}_{\mathcal{M}}$, then we can define a local parameterization $\omega_{\mathcal{M}}(K)$ of $\mathcal{C}_{\mathcal{M}}$ near $(\underline{\omega},\underline{K})$. We know [8] that

$$\pi_{\mathcal{M}}(\underline{\omega},\underline{K})A_j\pi_{\mathcal{M}}(\underline{\omega},\underline{K})=-\partial_j\omega_{\mathcal{M}}(\underline{K})\pi_{\mathcal{M}}(\underline{\omega},\underline{K}),\qquad j=1,2,3.$$

As $(\omega_l,-k_l)\in\mathcal{C}_{\mathcal{L}}$, it is easy to see that $(\omega_l,(0,0,-k_l))\in\mathcal{C}_{\mathcal{M}}$. Moreover, since $(\omega_l,-k_l)$ is a smooth point of $\mathcal{C}_{\mathcal{L}}$, $(\omega_l,(0,0,-k_l))$ is a smooth point of $\mathcal{C}_{\mathcal{M}}$ thanks to

Assumption 2.3. Thanks to the same assumption, a local parameterization may be used to write $\omega_{\mathcal{M}}(K) = \omega(|K|)$, where $\omega(\cdot)$ is a local parameterization of $\mathcal{C}_{\mathcal{L}}$ near $(\omega_l, -k_l)$.

Taking $\underline{K} = (0, 0, -k_l)$, one therefore has

$$\pi_{\mathcal{M}}(\omega_l, (0,0,-k_l))A_j\pi_{\mathcal{M}}(\omega_l, (0,0,-k_l)) = -\frac{\underline{K}_j}{k_l}\omega'(k_l).$$

Since $\underline{K}_1 = \underline{K}_2 = 0$, $\underline{K}_3 = -k_l$, and $\pi_M(\omega_l, \underline{K}) = \pi(\omega_l, -k_l)$, we obtain

$$\begin{aligned}
&\pi(\omega_l, -k_l)A_1\pi(\omega_l, -k_l) = 0,\\
&\pi(\omega_l, -k_l)A_2\pi(\omega_l, -k_l) = 0,\\
&\pi(\omega_l, -k_l)A_3\pi(\omega_l, -k_l) = \omega'(k_l),
\end{aligned}$$

which proves point (i).

The same proof shows that $\pi(\omega,k)A_1\pi(\omega,k) = \pi(\omega,k)A_2\pi(\omega,k) = 0$ and similarly $\pi(\omega,k)A_3\pi(\omega,k) = -\omega'(k)$ for every smooth point (ω,k) of $\mathcal{C}_{\mathcal{L}}$. Since we know by Assumption 2.1 that the set of the singular points of $\mathcal{C}_{\mathcal{L}}$ is discrete and hence has zero measure for $\mathcal{F}\mathcal{V}_{II}$ (because the spectrum of $\mathcal{V}_{II}$ is purely continuous), we can conclude that

$$\pi(D_{t_0,z_0})A_1\pi(D_{t_0,z_0})\mathcal{V}_{II} = \pi(D_{t_0,z_0})A_2\pi(D_{t_0,z_0})\mathcal{V}_{II} = 0$$

and that $\pi(D_{t_0,z_0})A_3\pi(D_{t_0,z_0})\mathcal{V}_{II} = -\omega'(D_{z_0})\mathcal{V}_{II}$, which proves point (iii).

For (ii), we still write $(0,0,-k_l) = \underline{K}$. Thanks to [8], we know that

$$\begin{aligned}
&\pi_{\mathcal{M}}(\omega_l, \underline{K})A_i\mathcal{M}^{-1}(\omega_l, \underline{K})A_j\pi_{\mathcal{M}}(\omega_l, \underline{K}) = \frac{1}{2}\pi_{\mathcal{M}}(\omega_l, \underline{K})\partial_{ij}^2\omega_{\mathcal{M}}(\underline{K})\\
&= \frac{1}{2}\pi_{\mathcal{M}}(\omega_l, \underline{K})\left(-\frac{\underline{K}_i\underline{K}_j}{|\underline{K}|^3}\omega'(|\underline{K}|) + \frac{\underline{K}_i\underline{K}_j}{|\underline{K}|^2}\omega''(|\underline{K}|) + \frac{\delta_{ij}}{|\underline{K}|}\omega'(|\underline{K}|)\right),
\end{aligned}$$

where δ_{ij} denotes Kronecker's symbol, $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

Since $\underline{K}_1 = \underline{K}_2 = 0$, one has

$$\begin{aligned}
&\pi_{\mathcal{M}}(\omega_l, \underline{K})A_1\mathcal{M}^{-1}(\omega_l, \underline{K})A_2\pi_{\mathcal{M}}(\omega_l, \underline{K}) = 0,\\
&\pi_{\mathcal{M}}(\omega_l, \underline{K})A_2\mathcal{M}^{-1}(\omega_l, \underline{K})A_1\pi_{\mathcal{M}}(\omega_l, \underline{K}) = 0,
\end{aligned}$$

and since $\underline{K}_3 = -k_l$,

$$\begin{aligned}
\pi_{\mathcal{M}}A_1\mathcal{M}^{-1}(\omega_l, \underline{K})A_1\pi_{\mathcal{M}} &= \pi_{\mathcal{M}}A_2\mathcal{M}^{-1}(\omega_l, \underline{K})A_2\pi_{\mathcal{M}}\\
&= \frac{\omega'(k_l)}{2k_l}\pi_{\mathcal{M}}(\omega_l, \underline{K})
\end{aligned}$$

and $\pi_{\mathcal{M}}A_3\mathcal{M}^{-1}(\omega_l, \underline{K})A_3\pi_{\mathcal{M}} = \frac{\omega''(k_l)}{2}\pi_{\mathcal{M}}$.

Since $\pi_{\mathcal{M}}(\omega_l, \underline{K}) = \pi(\omega_l, -k_l)$ and $\mathcal{M}^{-1}(\omega_l, \underline{K}) = \mathcal{L}^{-1}(\omega_l, -k_l)$ we obtain (ii).

The same reasoning as in (iii) yields (iv). ∎

Therefore, according to Proposition 2.4 and systems (2.20), (2.21), the leading term $\mathcal{U}_0(\tau, T, X, Y, Z, t_0, z_0) = \mathcal{U}_{0,I,1}(\tau, T, X, Y, Z)e^{i\theta} + \text{c.c.} + \mathcal{U}_{0,II}(\tau, T, X, Y, Z, t_0, z_0)$ is found by solving (if possible)

$$\left\{\begin{aligned}
&\pi(\omega_l, -k_l)\mathcal{U}_{0,I,1} = \mathcal{U}_{0,I,1},\\
&(\partial_T + \omega'(k_l)\partial_Z)\mathcal{U}_{0,I,1} = 0,\\
&\partial_\tau\mathcal{U}_{0,I,1} + i\frac{\omega'(k_l)}{2k_l}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,I,1} + i\frac{\omega''(k_l)}{2}\partial_Z^2\mathcal{U}_{0,I,1}\\
&\qquad\qquad +\pi(\omega_l, -k_l)f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}}) = 0
\end{aligned}\right. \tag{2.22}$$

and

$$
(2.23)\quad \begin{cases} \pi(D_{t_0,z_0})\mathcal{U}_{0,II} = \mathcal{U}_{0,II}, \\ (\partial_T - \omega'(D_{z_0})\partial_Z)\mathcal{U}_{0,II} = 0, \\ \partial_\tau \mathcal{U}_{0,II} + (\partial_T - \omega'(D_{z_0})\partial_Z)\pi(D_{z_0})\mathcal{U}_{1,II} + i\dfrac{\omega'(D_{z_0})}{2D_{z_0}}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,II} \\ \qquad\qquad +i\dfrac{\omega''(D_{z_0})}{2}\partial_Z^2\mathcal{U}_{0,II} + \psi^\delta(D_{z_0})\pi(D_{t_0,z_0})[f(\mathcal{U}_0)]_{II} = 0. \end{cases}
$$

Remark 2.6. (i) The notation $\frac{\omega'(D_{z_0})}{2D_{z_0}}$ is ambiguous since one could think that this Fourier multiplier does not depend on D_{t_0}. In fact, for any $k \in \mathbb{R}$, there are several ω_j such that $(\omega_j, k) \in \mathcal{C}_\mathcal{L}$. In Fourier variables, $\omega'(D_{z_0})$ therefore reads $\mathcal{F}\omega'(D_{z_0})(\omega, k) = \omega_j'(k)$, where the subscript j is such that $(\omega_j, k) = (\omega, k)$ and thus depends on w.

(ii) In general $\frac{\omega'(D_{z_0})}{2D_{z_0}}$ is not a Fourier multiplier of $E^s_{\tau^*}$. However, it is well defined here since it is applied to $\mathcal{U}_{0,II}$, which satisfies Assumption 2.2.

2.2. Analysis of the nonlinearity. We have to work more on these profile equations before trying to solve them. In particular, it is essential to simplify the nonlinearity $[f(\mathcal{U}_0)]_{II}$ which appears in (2.23). In this section, we show a striking result. *The nonlinearity* $[f(\mathcal{U}_0)]_{II}$ *is in fact linear.*

We recall that there exists a trilinear mapping F such that $F(u,u,u) = f(u)$ for all $u \in C^n$. The nonlinearity $f(\mathcal{U}_0)_{II}$ which appears in the evolution equation of $\mathcal{U}_{0,II}$ therefore reads

$$
\begin{aligned}
&F(\mathcal{U}_{0,I} + \mathcal{U}_{0,II}, \mathcal{U}_{0,I} + \mathcal{U}_{0,II}, \mathcal{U}_{0,I} + \mathcal{U}_{0,II})]_{II} = F(\mathcal{U}_{0,II}, \mathcal{U}_{0,II}, \mathcal{U}_{0,II}) \\
&+ F(\mathcal{U}_{0,I}, \mathcal{U}_{0,II}, \mathcal{U}_{0,II}) + F(\mathcal{U}_{0,II}, \mathcal{U}_{0,I}, \mathcal{U}_{0,II}) + F(\mathcal{U}_{0,II}, \mathcal{U}_{0,II}, \mathcal{U}_{0,I}) \\
&+ \big(F(\mathcal{U}_{0,I,1}, \mathcal{U}_{0,I,1}, \mathcal{U}_{0,II}) + F(\mathcal{U}_{0,II}, \mathcal{U}_{0,I,1}, \mathcal{U}_{0,I,1}) + F(\mathcal{U}_{0,I,1}, \mathcal{U}_{0,II}, \mathcal{U}_{0,I,1})\big)e^{2i\theta} \\
&+ \big(F(\overline{\mathcal{U}_{0,I,1}}, \overline{\mathcal{U}_{0,I,1}}, \mathcal{U}_{0,II}) + F(\mathcal{U}_{0,II}, \overline{\mathcal{U}_{0,I,1}}, \overline{\mathcal{U}_{0,I,1}}) + F(\overline{\mathcal{U}_{0,I,1}}, \mathcal{U}_{0,II}, \overline{\mathcal{U}_{0,I,1}})\big)e^{-2i\theta} \\
&+ F(\mathcal{U}_{0,I,1}, \overline{\mathcal{U}_{0,I,1}}, \mathcal{U}_{0,II}) + F(\overline{\mathcal{U}_{0,I,1}}, \mathcal{U}_{0,I,1}, \mathcal{U}_{0,II}) + F(\mathcal{U}_{0,I,1}, \mathcal{U}_{0,II}, \overline{\mathcal{U}_{0,I,1}}) \\
&+ F(\overline{\mathcal{U}_{0,I,1}}, \mathcal{U}_{0,II}, \mathcal{U}_{0,I,1}) + F(\mathcal{U}_{0,II}, \mathcal{U}_{0,I,1}, \overline{\mathcal{U}_{0,I,1}}) + F(\mathcal{U}_{0,II}, \overline{\mathcal{U}_{0,I,1}}, \mathcal{U}_{0,I,1}).
\end{aligned}
$$

This expression can be simplified a lot because the role of the components with a purely continuous spectrum in the nonlinearity is not often efficient. This is the object of the following lemma.

LEMMA 2.5. *Suppose Assumption* 2.1 *is satisfied and let* $\mathcal{V}_{II} \in E^s_{\tau^*}$ *be a profile with purely continuous spectrum and such that* $\mathrm{Sp}\ \mathcal{V}_{II} \subset \mathcal{C}_\mathcal{L}$. *Take also* $a, b \in \mathbb{C}^n$. *Then one has*

(i) $\pi(D_{t_0,z_0})F(\mathcal{V}_{II}, \mathcal{V}_{II}, \mathcal{V}_{II}) = 0$;
(ii) $\pi(D_{t_0,z_0})F(ae^{i\theta}, \mathcal{V}_{II}, \mathcal{V}_{II}) = \pi(D_{t_0,z_0})F(ae^{-i\theta}, \mathcal{V}_{II}, \mathcal{V}_{II}) = 0$;
(iii) $\pi(D_{t_0,z_0})F(ae^{i\theta}, be^{i\theta}, \mathcal{V}_{II}) = \pi(D_{t_0,z_0})F(ae^{-i\theta}, be^{-i\theta}, \mathcal{V}_{II}) = 0$.

Proof. Let $\mathcal{V}_{II} \in E^s_{\tau^*}$ be as in the lemma. For (τ, T) fixed, we introduce $\mu := \mathcal{F}\mathcal{V}_{II}(\tau, T)$ and denote by $v(\mu)$ the total variation of μ. Thanks to the Radon–Nikodým property, we can write, for all Borel sets E of $\mathbb{R}^2$,

$$
\mu(E) = \int_E r_\mu(\xi) v(\mu)(d\xi),
$$

where r_μ is an $H^s(\mathbb{R}^3)^n$-valued integrable function such that $||r_\mu(\xi)||_{H^s} = 1$ for $v(\mu)$ for almost all ξ.

Introducing $\nu := \mathcal{F}(\pi(D_{t_0,z_0})F(\mathcal{V}_{II}, \mathcal{V}_{II}, \mathcal{V}_{II}))$, the first point of the lemma will be proved if we can show that $\nu = 0$, i.e., that $v(\nu)(\mathbb{R}^2) = 0$. One has

$$v(\nu)(\mathbb{R}^2) = \int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} ||\pi(\xi_1+\xi_2+\xi_3)F(r_\mu(\xi_1), r_\mu(\xi_2), r_\mu(\xi_3))||_{H^s(\mathbb{R}^3)^n} \times v(\mu)(d\xi_1)v(\mu)(d\xi_2)v(\mu)(d\xi_3).$$

Since Sp $\mathcal{V}_{II} \subset \mathcal{C}_\mathcal{L}$, one can take $r_\mu(\xi) = 0$ if $\xi \notin \mathcal{C}_\mathcal{L}$. Hence,

$$\begin{aligned} v(\nu)(\mathbb{R}^2) &= \int_{\mathcal{C}_\mathcal{L}}\int_{\mathcal{C}_\mathcal{L}}\int_{\mathcal{C}_\mathcal{L}} ||\pi(\xi_1+\xi_2+\xi_3)F(r_\mu(\xi_1), r_\mu(\xi_2), r_\mu(\xi_3))||_{H^s(\mathbb{R}^3)^n} \\ &\qquad \times v(\mu)(d\xi_1)v(\mu)(d\xi_2)v(\mu)(d\xi_3) \\ &= \int_{\mathcal{C}_\mathcal{L}}\int_{\mathcal{C}_\mathcal{L}}\left[\int_{\mathcal{C}_\mathcal{L}} ||\pi(\xi_1+\xi_2+\xi_3)F(r_\mu(\xi_1), r_\mu(\xi_2), r_\mu(\xi_3))||_{H^s(\mathbb{R}^3)^n} v(\mu)(d\xi_3)\right] \\ &\qquad \times v(\mu)(d\xi_1)v(\mu)(d\xi_2). \end{aligned}$$

Notice that if $\xi_1+\xi_2 \neq 0$, then the set of $\xi \in \mathcal{C}_\mathcal{L}$ such that $\xi_1+\xi_2+\xi \in \mathcal{C}_\mathcal{L}$ is discrete. Indeed following [6] and [5], the set of $\xi \in \mathcal{C}_\mathcal{L}$ such that $\xi_1+\xi_2+\xi \in \mathcal{C}_\mathcal{L}$ is either a sheet of $\mathcal{C}_\mathcal{L}$ or an algebraic submanifold of $\mathcal{C}_\mathcal{L}$ of strictly lower dimension. Since $\mathcal{C}_\mathcal{L}$ is an algebraic manifold of dimension one, the set of $\xi \in \mathcal{C}_\mathcal{L}$ such that $\xi_1+\xi_2+\xi \in \mathcal{C}_\mathcal{L}$ is then either a sheet of $\mathcal{C}_\mathcal{L}$ or a discrete set. The first case is not possible thanks to Assumption 2.1 because we would have two parallel sheets of $\mathcal{C}_\mathcal{L}$. Therefore, the set of points $\xi \in \mathcal{C}_\mathcal{L}$ such that $\xi_1+\xi_2+\xi \in \mathcal{C}_\mathcal{L}$ is discrete if $\xi_1+\xi_2 \neq 0$. Since $v(\mu)(\{\xi\}) = 0$ for all $\xi \in \mathbb{R}^2$, we can conclude that

$$\int_{\mathcal{C}_\mathcal{L}} ||\pi(\xi_1+\xi_2+\xi_3)F(r_\mu(\xi_1), r_\mu(\xi_2), r_\mu(\xi_3))||_{H^s(\mathbb{R}^3)^n} v(\mu)(d\xi_3) = 0,$$

when $\xi_1+\xi_2 \neq 0$.

Therefore, one has

$$v(\nu)(\mathbb{R}^2) = \int_{\mathcal{C}_\mathcal{L}}\int_{\mathcal{C}_\mathcal{L}} ||\pi(\xi_3)F(r_\mu(\xi_1), r_\mu(-\xi_1), r_\mu(\xi_3))||_{H^s(\mathbb{R}^3)^n} v(\mu)(\{-\xi_1\}) \times v(\mu)(d\xi_1)v(\mu)(d\xi_3).$$

Since $v(\mu)(\{\xi\}) = 0$ for all $\xi \in \mathbb{R}^2$, the above quantity is equal to 0, i.e., $v(\nu)(\mathbb{R}^2) = 0$, which proves point (i).

For (ii), introduce $\lambda := \mathcal{F}(\pi(D_{t_0,z_0})F(ae^{i\theta}, \mathcal{V}_{II}, \mathcal{V}_{II}))$. One has

$$v(\lambda)(\mathbb{R}^2) = \int_{\mathcal{C}_\mathcal{L}}\int_{\mathcal{C}_\mathcal{L}} ||\pi((\omega_l, -k_l)+\xi_1+\xi_2)F(a, r_\mu(\xi_1), r_\mu(\xi_2))||_{H^s(\mathbb{R}^3)^n} \times v(\mu)(d\xi_1)v(\mu)(d\xi_2).$$

With the same reasoning as in (i), one can prove that this expression vanishes, thus yielding point (ii).

Point (iii) is a direct consequence of Assumption 2.1 and, more precisely, of the fact that two different sheets of $\mathcal{C}_\mathcal{L}$ are never parallel. □

Remark 2.7. Lemma 2.5 can easily be generalized to N-linear nonlinear functions F. When $\mathcal{V}_{II}$ appears twice or more in the arguments of F, the term vanishes. When it appears once only, the spectrum of this nonlinear term is a translation of the spectrum of $\mathcal{V}_{II}$, and hence this nonlinear term vanishes unless the intersection of this spectrum with $\mathcal{C}_\mathcal{L}$ has the same dimension as $\mathcal{C}_\mathcal{L}$.

Thanks to Lemma 2.5, many terms vanish when we apply the operator $\pi(D_{t_0,z_0})$ to the nonlinearity. One then finds $\pi(D_{t_0,z_0})[f(\mathcal{U}_0)]_{II} = \pi(D_{t_0,z_0})F^S(\mathcal{U}_{0,I,1},\overline{\mathcal{U}_{0,I,1}},\mathcal{U}_{0,II})$, where the symmetrized function F^S associated to F is defined as

$$F^S(a,b,c) = F(a,b,c) + F(a,c,b) + F(b,a,c) + F(c,a,b) + F(b,c,a) + F(c,b,a)$$

for all $a,b,c \in \mathbb{C}^n$.

The equations for the profile $\mathcal{U}_{0,II}$ are thus

$$\text{(2.24)}\quad \begin{cases} \pi(D_{t_0,z_0})\mathcal{U}_{0,II} = \mathcal{U}_{0,II}, \\ (\partial_T - \omega'(D_{z_0})\partial_Z)\mathcal{U}_{0,II} = 0, \\ \partial_\tau \mathcal{U}_{0,II} + (\partial_T - \omega'(D_{z_0})\partial_Z)\mathcal{U}_{1,II} + i\dfrac{\omega'(D_{z_0})}{2D_{z_0}}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,II} \\ \qquad +i\frac{\omega''(D_{z_0})}{2}\partial_Z^2\mathcal{U}_{0,II} + \psi^\delta(D_{z_0})\pi(D_{t_0,z_0})F^S(\mathcal{U}_{0,I,1},\overline{\mathcal{U}_{0,I,1}},\mathcal{U}_{0,II}) = 0, \end{cases}$$

and are therefore coupled with the equations found above for the amplitude $\mathcal{U}_{0,I,1}$,

$$\text{(2.25)}\quad \begin{cases} \pi(\omega_l,-k_l)\mathcal{U}_{0,I,1} = \mathcal{U}_{0,I,1}, \\ (\partial_T + \omega'(k_l)\partial_Z)\mathcal{U}_{0,I,1} = 0, \\ \partial_\tau \mathcal{U}_{0,I,1} + i\dfrac{\omega'(k_l)}{2k_l}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,I,1} + i\dfrac{\omega''(k_l)}{2}\partial_Z^2\mathcal{U}_{0,I,1} \\ \qquad +\pi(\omega_l,-k_l)f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}}) = 0. \end{cases}$$

Remark 2.8. One can see that the evolution equation for the oscillating mode is the usual *uncoupled* NLS equation. The evolution equation of $\mathcal{U}_{0,II}$ is in turn *linear* but *coupled* with $\mathcal{U}_{0,I,1}$. The next step is to prove that this coupling is not efficient. This is the object of the next section.

2.3. Solving the profile equations in absence of low frequencies. If system (2.25) can easily be solved by standard Picard iterates (see [8], for instance), this is not the case for system (2.24), which deals with the continuous spectrum component of $\mathcal{U}_0$. Moreover, one can see that it is not possible to take $\pi(D_{t_0,z_0})\mathcal{U}_{1,II} = 0$ as for the discrete spectrum component, since the term $F^S(\mathcal{U}_{0,I,1},\overline{\mathcal{U}_{0,I,1}},\mathcal{U}_{0,II})$ makes the last equation of (2.24) obviously incompatible with the transport equation $(\partial_T - \omega'(D_{z_0})\partial_Z)\mathcal{U}_{0,II} = 0$.

In fact, as in the papers where various group velocities are studied [10], [11], only a good choice of $\pi(D_{t_0,z_0})\mathcal{U}_{1,II} = 0$ can allow the solubility for (2.24). Inspired by the method and arguments of [10] and [11] (i.e., interactions between components traveling at different speeds do not affect the main profile), we decompose (2.24) as follows:

$$\text{(2.26)}\quad \begin{cases} \pi(D_{t_0,z_0})\mathcal{U}_{0,II} = \mathcal{U}_{0,II}, \\ (\partial_T - \omega'(D_{z_0})\partial_Z)\mathcal{U}_{0,II} = 0, \\ \partial_\tau \mathcal{U}_{0,II} + i\dfrac{\omega'(D_{z_0})}{2D_{z_0}}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,II} + i\dfrac{\omega''(D_{z_0})}{2}\partial_Z^2\mathcal{U}_{0,II} = 0, \\ (\partial_T - \omega'(D_{z_0})\partial_Z)\mathcal{U}_{1,II} = -\psi^\delta(D_{z_0})\pi(D_{t_0,z_0})F^S(\mathcal{U}_{0,I,1},\overline{\mathcal{U}_{0,I,1}},\mathcal{U}_{0,II}). \end{cases}$$

We can now state a solubility result, provided that the function $\mathbf{U}_{II}^0$ has no low frequencies.

PROPOSITION 2.6. *Let $\sigma \geq s$, and let $R > 0$ be such that $\mathbf{U}^0 = \mathbf{U}^0_I + \mathbf{U}^0_{II} \in A^\sigma_0$ and $\|\mathbf{U}^0\|_{A^\sigma_0} \leq R$. Suppose, moreover, that $\mathbf{U}^0_I = \mathbf{U}^0_{I,1} e^{i\theta} + c.c.$ and*

$$\pi(\omega_l, -k_l)\mathbf{U}^0_{I,1} = \mathbf{U}^0_{I,1}, \qquad \pi(D_{t_0,z_0})\mathbf{U}^0_{II} = \mathbf{U}^0_{II},$$

and that there exists $\delta > 0$ such that Sp $\mathbf{U}^0_{II} \subset \{(\omega, k), |k| > \delta\}$.

Then there exists $\tau_2^* > 0$, which depends on R but not on ε nor on δ, such that there exists

- *a unique $\mathcal{U}_{0,I,1} = \pi(\omega_l, -k_l)\mathcal{U}_{0,I,1} \in \mathcal{C}_b([0,\tau_2^*] \times \mathbb{R}_T, H^\sigma(\mathbb{R}^3)^n)$ solving*

$$\begin{cases} (\partial_T + \omega'(k_l)\partial_Z)\mathcal{U}_{0,I,1} = 0, \\ \partial_\tau \mathcal{U}_{0,I,1} + i\dfrac{\omega'(k_l)}{2k_l}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,I,1} + i\dfrac{\omega''(k_l)}{2}\partial_Z^2 \mathcal{U}_{0,I,1} \\ \qquad\qquad + \pi(\omega_l, -k_l) f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}}) = 0, \\ \mathcal{U}_{0,I,1}|_{\tau=T=0} = \mathbf{U}^0_{I,1}; \end{cases} \tag{2.27}$$

- *a unique $\mathcal{U}_{0,II} = \pi(D_{t_0,z_0})\mathcal{U}_{0,II} \in A^\sigma_{\tau_2^*}$ solving*

$$\begin{cases} (\partial_T - \omega'(D_{z_0})\partial_Z)\mathcal{U}_{0,II} = 0, \\ \partial_\tau \mathcal{U}_{0,II} + i\dfrac{\omega'(D_{z_0})}{2D_{z_0}}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,II} + i\dfrac{\omega''(D_{z_0})}{2}\partial_Z^2 \mathcal{U}_{0,II} = 0, \\ \mathcal{U}_{0,II}|_{\tau=T=0} = \mathbf{U}^0_{II}; \end{cases} \tag{2.28}$$

- *a unique $\mathcal{U}_{1,II} \in E^\sigma_{\tau_2^*}$ solving*

$$\begin{cases} (Id - \pi(D_{t_0,z_0}))\mathcal{U}_{1,II} = i\mathcal{L}^{-1}(D_{t_0,z_0})A(\partial_{X,Y,Z})\mathcal{U}_{0,II}, \\ \psi^\delta(D_{z_0})\pi(D_{t_0,z_0})\mathcal{U}_{1,II} = \pi(D_{t_0,z_0})\mathcal{U}_{1,II}, \\ (\partial_T - \omega'(D_{z_0})\partial_Z)\pi(D_{z_0,t_0})\mathcal{U}_{1,II} \\ \qquad\qquad = -\psi^\delta(D_{z_0})\pi(D_{t_0,z_0})F^S(\mathcal{U}_{0,I,1}, \overline{\mathcal{U}_{0,I,1}}, \mathcal{U}_{0,II}), \\ \mathcal{U}_{1,II}|_{\tau=T=0} = 0. \end{cases} \tag{2.29}$$

Moreover $\mathcal{U}_{0,II}$ satisfies Assumption 2.2 *with the same δ as above, and we have the upper bound $\|\mathcal{U}_0\|_{A^\sigma_{\tau_2^*}} \leq 2R$.*

Proof. The proof of the existence and the uniqueness of the solution of (2.27) is done by standard Picard iterates, as in [8], for instance. Since $\mathcal{U}_{0,I,1}$ must solve the transport equation $(\partial_T + \omega'(k_l)\partial_Z)\mathcal{U}_{0,I,1} = 0$, there exists $\mathtt{U}_{0,I,1} \in \mathcal{C}([0,\tau_2^*], H^s(\mathbb{R}^3)^n)$ such that

$$\mathcal{U}_{0,I,1}(\tau, T, X, Y, Z) = \mathtt{U}_{0,I,1}(\tau, Z - \omega'(k_l)T, X, Y). \tag{2.30}$$

Moreover, the Schrödinger equation that $\mathcal{U}_{0,I,1}$ must satisfy implies, for $\mathtt{U}_{0,I,1}$,

$$\partial_\tau \mathtt{U}_{0,I,1} + i\frac{\omega'(k_l)}{2k_l}(\partial_X^2 + \partial_Y^2)\mathtt{U}_{0,I,1} + i\frac{\omega''(k_l)}{2}\partial_Z^2 \mathtt{U}_{0,I,1} + \pi(\omega_l, -k_l) f'(\mathtt{U}_{0,I,1})(\overline{\mathtt{U}_{0,I,1}}) = 0.$$

Existence and uniqueness of such a $\mathtt{U}_{0,I,1} \in \mathcal{C}([0,\tau_2^*], H^s(\mathbb{R}^3)^n)$ can be proved by standard Picard iterates, and we thus obtain point (i) of the theorem.

In order to prove (ii), introduce $\lambda := \mathcal{F}\mathcal{U}_{0,II}$ and $\lambda_0 := \mathcal{F}\mathbf{U}^0_{II}$. There is a unique solution of (2.28) in the sense of distributions, given by

$$\widehat{\lambda}(\tau, T) = e^{-i\tau(\frac{\omega'(k)}{2k}(\eta_1^2 + \eta_2^2) + \frac{\omega''(k)}{2}\eta_3^2)} e^{iT\omega'(k)\eta_3}\widehat{\lambda}_0, \tag{2.31}$$

where $\widehat{\ }$ denotes the Fourier transform with respect to the variables (X, Y, Z), and the measure $\widehat{\lambda}$ is defined for all Borel sets E of $\mathbb{R}^2$ by $\widehat{\lambda}(E) := \widehat{\lambda(E)}$.

The distribution λ defined above is in $\mathcal{C}_b([0,\tau_2^*]\times\mathbb{R}_T, \mathcal{BV}(\mathbb{R}^2_\xi, H^\sigma(\mathbb{R}^3)^n))$, as one can prove with the same arguments as in Theorem 4 of [12], which proves point (ii) of the theorem.

By the explicit expression (2.31) and under the assumption made on $\mathbf{U}^0_{II}$, we also know that $\mathcal{U}_{0,II}$ satisfies Assumption 2.2 and that $||\mathcal{U}_{0,II}||_{A^\sigma_{\tau_2^*}} \leq \mathbb{R}$.

Before solving the next equation, notice that $\pi(D_{t_0,z_0})F^S(\mathcal{U}_{0,I,1}, \overline{\mathcal{U}_{0,I,1}}, \mathcal{U}_{0,II}) \in A^\sigma_{\tau_2^*}$, thanks to the above results and to Proposition 1.3 . We denote by μ the Fourier transform of this profile and also denote by ν the Fourier transform of the profile $\mathcal{U}_{1,II}$. The last equation of system (2.29) has a unique solution in the sense of the distributions, which reads

$$\widehat{\nu}(\tau, T) = -\int_0^T e^{i(T-u)\omega'(k)}\eta_3\psi^\delta(k)\widehat{\mu}(\tau,u)du. \tag{2.32}$$

Here again, we can prove that ν is in $\mathcal{C}([0,\tau_2^*]\times\mathbb{R}_T, \mathcal{BV}(\mathbb{R}^2_\xi, H^\sigma(\mathbb{R}^3)^n))$ so that $\mathcal{U}_{1,II} \in E^\sigma_{\tau_2^*}$. From this expression it is also clear that $\mathrm{Sp}\,\mathcal{U}_{1,II} \subset \{(\omega,k), |k| > \delta\}$, i.e., that the first equation of (2.29) is also satisfied. □

Remark 2.9. (i) The profile equations are solved in $E^\sigma_{\tau_2^*}$ for all $\sigma \geq s$, not only in $E^s_{\tau_2^*}$, because if one wants the residual term $\mathcal{R}^\varepsilon$ to be in $E^s_{\tau_2^*}$, $\mathcal{U}_0$ must be in $E^{s+4}_{\tau_2^*}$, as we will see in the next section.

(ii) The remaining profiles $(Id - \pi(\omega_l, -k_l))\mathcal{U}_{1,I,1}$, $(Id - \pi(D_{t_0,z_0}))\mathcal{U}_{1,II}$, and $\mathcal{U}_2$ can be found in terms of the profiles given by the above theorem, thanks to the expressions given by (2.4)–(2.5), (2.12), (2.16), and (2.18).

2.4. Stability in the absence of low frequencies. In this part, we want to prove that the solution of diffractive optics gives a good approximation of the exact solution of (1.1), provided that low frequencies are excluded. To get this result, we prove that the residual associated to the approximate solution is small and that the approximate solution is close to the exact solution $\mathbf{u}^\varepsilon$ of (1.1).

Suppose $\mathcal{U}_0$ and $\pi(D_{t_0,z_0})\mathcal{U}_{1,II}$ are given by Proposition 2.6 with initial condition $\mathbf{U}^0$ without frequencies lower than $\delta > 0$. $\mathcal{U}_{0,II}$ then satisfies Assumption 2.2, $\pi(D_{t_0,z_0})\mathcal{U}_{1,II}$ fulfills condition (2.13), and $\mathcal{U}_{2,I}$ and $\mathcal{U}_{2,II}$ can then be constructed with the results of section 2.1. Before proving that $\mathcal{U}^\varepsilon = \mathcal{U}_0 + \varepsilon\mathcal{U}_1 + \varepsilon^2\mathcal{U}_2$ is an approximate solution of the singular equation (1.5) in the sense that the residual remains small, we need estimates of the profiles $\mathcal{U}_j$, $j = 0, 1, 2$, and of the residual terms $\mathcal{R}_j$, $j \geq 2$.

LEMMA 2.7. *Let $\sigma \geq s+4$, $\sigma' \geq s$, and $\delta \in (0,1)$ and suppose $\mathcal{U}_0 \in A^\sigma_{\tau_2^*}$ and $\pi(D_{t_0,z_0})\mathcal{U}_{1,II} \in E^\sigma_{\tau_2^*}$ are given by Proposition* 2.6. *Take $\mathcal{U}_{1,I}$, $(Id - \pi(D_{t_0,z_0}))\mathcal{U}_{1,II}$ and $\mathcal{U}_2$ as computed in section* 2.1. *Then*

(i) *the two components $\mathcal{U}_{1,I}$ and $\mathcal{U}_{1,II}$ of $\mathcal{U}_1$ are controlled as follows:*

$$\|\mathcal{U}_{1,I}\|_{A^{\sigma'}_{\tau_2^*}} \leq C\|\mathcal{U}_{0,I}\|_{A^{\sigma'+1}_{\tau_2^*}}, \quad \|(Id - \pi(D_{t_0,z_0}))\mathcal{U}_{1,II}\|_{A^{\sigma'}_{\tau_2^*}} \leq \frac{C}{\delta}\|\mathcal{U}_{0,II}\|_{A^{\sigma'+1}_{\tau_2^*}},$$

and

$$\|\pi(D_{t_0,z_0})\mathcal{U}_{1,II}(T)\|_{E^{\sigma'}_{\tau_2^*}} \leq CT\|\mathcal{U}_{0,I}\|^2_{A^{\sigma'}_{\tau_2^*}}\|\mathcal{U}_{0,II}\|_{A^{\sigma'}_{\tau_2^*}} \qquad \forall T \geq 0;$$

(ii) *the discrete spectrum component* $\mathcal{U}_{2,I}$ *of* $\mathcal{U}_2$ *is controlled as follows:*

$$\|\mathcal{U}_{2,I}\|_{A^{\sigma'}_{\tau_2^*}} \leq C\left(\|\mathcal{U}_{0,I}\|_{A^{\sigma'+2}_{\tau_2^*}} + \|\mathcal{U}_{0,I}\|^3_{A^{\sigma'}_{\tau_2^*}}\right),$$

while $\mathcal{U}_{2,II}$ *satisfies*

$$\|\mathcal{U}_{2,II}\|_{A^{\sigma'}_{\tau_2^*}} \leq \frac{C}{\delta}\left(\frac{1}{\delta}\|\mathcal{U}_{0,II}\|_{A^{\sigma'+2}_{\tau_2^*}} + \|\mathcal{U}_0\|^3_{A^{\sigma'}_{\tau_2^*}} + T\|\mathcal{U}_{0,I}\|^2_{A^{\sigma'+1}_{\tau_2^*}}\|\mathcal{U}_{0,II}\|_{A^{\sigma'+1}_{\tau_2^*}}\right);$$

(iii) *rough estimates of the profiles* $\mathcal{R}_{j\geq 2}$ *are given by*

$$\|\mathcal{R}_{j,I}\|_{A^{\sigma'}_{\tau_2^*}} \leq h_1\left(\|\mathcal{U}_{0,I}\|_{A^{\sigma'+4}_{\tau_2^*}}\right) \quad \text{and} \quad \|\mathcal{R}_{j,II}(T)\|_{E^{\sigma'}_{\tau_2^*}} \leq \frac{T}{\delta^6}h_2\left(\|\mathcal{U}_0\|_{A^{\sigma'+4}_{\tau_2^*}}\right) \quad \forall T \geq 0,$$

where h_1 *and* h_2 *are smooth positive functions defined on* $\mathbb{R}^+$ *and independent of* $\delta \in (0,1)$ *and of* $T \geq 0$.

Proof. The estimate of $\mathcal{U}_{1,I}$ is easily deduced from (2.4) and (2.6), while Lemma 2.1 and (2.5)–(2.6) yield the estimate of $(Id - \pi(D_{t_0,z_0}))\mathcal{U}_{1,II}$. The estimate of $\pi(D_{t_0,z_0})\mathcal{U}_{1,II}$ is a consequence of (2.32).

The modes ± 3 of $\mathcal{U}_{2,I}$ are controlled using (2.10) together with the algebra properties of $A^{\sigma'}_{\tau_2^*}$. The modes ± 1 are controlled as stated in the lemma as a consequence of (2.12) and (2.18). Lemma 2.1 and (2.16), (2.18) are used to estimate $\mathcal{U}_{2,II}$.

The estimates of the profiles $\mathbb{R}_j$ follow directly from the explicit formulae of these profiles given in (1.8), from Lemma 2.1 and the equations satisfied by $\mathcal{U}_{1\leq j\leq 3}$, and from the estimates of these profiles computed above. ∎

PROPOSITION 2.8. *Let* $\sigma \geq s+4$ *and* $\delta \in (0,1)$ *and suppose* $\mathcal{U}_0$ *and* $\pi(D_{t_0,z_0})\mathcal{U}_{1,II}$ *are as given by Proposition* 2.6. *Take* $\mathcal{U}_{1,I}$, $(Id - \pi(D_{t_0,z_0}))\mathcal{U}_{1,II}$ *and* $\mathcal{U}_2$ *as computed in section* 2.1 *and let* $\mathcal{U}^\varepsilon = \mathcal{U}_0 + \varepsilon\mathcal{U}_1 + \varepsilon^2\mathcal{U}_2 \in E^{s+2}_{\tau_2^*} \subset E^s_{\tau_2^*}$.

Then the profile $\underline{\mathcal{U}}^\varepsilon$ *defined as* $\underline{\mathcal{U}}^\varepsilon(\tau, X, Y, Z, t_0, z_0) = \mathcal{U}^\varepsilon(\tau, \tau/\varepsilon, X, Y, Z, t_0, z_0)$ *is in* $B^{s+2}_{\tau_2^*}$ *and is bounded uniformly in* $\varepsilon \in (0,1)$.

If δ *is small enough,* $\underline{\mathcal{U}}^\varepsilon$ *is an approximate solution of the singular equation* (1.5). *More precisely, for any* $\mu > 0$ *there exists* $\delta(\mu) > 0$ *such that if* $0 < \delta < \delta(\mu)$, $\underline{\mathcal{U}}^\varepsilon$ *satisfies*

$$\limsup_{\varepsilon\to 0} \|\partial_\tau\underline{\mathcal{U}}^\varepsilon + \varepsilon^{-1}(A_1\partial_X + A_2\partial_Y + A_3\partial_Z)\underline{\mathcal{U}}^\varepsilon + \varepsilon^{-2}(\partial_{t_0} + A_3\partial_{z_0} + L_0)\underline{\mathcal{U}}^\varepsilon + f(\underline{\mathcal{U}}^\varepsilon)\| < \mu/3,$$

where the norm is taken in $B^s_{\tau_2^*}$.

Proof. Define, for $j = 1,2,3$, $\underline{\mathcal{U}}^\varepsilon_j(\tau, X, Y, Z, t_0, z_0) = \mathcal{U}_j(\tau, \tau/\varepsilon, X, Y, Z, t_0, z_0)$. One has $\underline{\mathcal{U}}^\varepsilon_0 \in B^{s+2}_{\tau_2^*}$ since $\mathcal{U}_0 \in A^{s+2}_{\tau_2^*}$. Similarly, $\underline{\mathcal{U}}^\varepsilon_{1,I}$ and $(Id - \pi(D_{t_0,z_0}))\underline{\mathcal{U}}^\varepsilon_{1,II}$ are in $B^{s+2}_{\tau_2^*}$. Their norm in this space is obviously uniformly bounded in $\varepsilon \in (0,1)$.

Since $\pi(D_{t_0,z_0})\mathcal{U}_{1,II} \notin A^{s+2}_{\tau_2^*}$, we cannot apply the same reasoning for this component. However, point (i) of Lemma 2.7 asserts that $\varepsilon\pi(D_{t_0,z_0})\mathcal{U}_{1,II}$ satisfies, for all $T \geq 0$,

$$\|\varepsilon\pi(D_{t_0,z_0})\mathcal{U}_{1,II}(T)\|_{E^{s+2}_{\tau_2^*}} \leq \varepsilon T\|\mathcal{U}_{0,I}\|^2_{A^{s+2}_{\tau_2^*}}\|\mathcal{U}_{0,II}\|_{A^{s+2}_{\tau_2^*}},$$

and since one has

$$\|\pi(D_{t_0,z_0})\underline{\mathcal{U}}^\varepsilon_{1,II}\|_{B^{s+2}_{\tau_2^*}} \leq \sup_{T\in[0,\tau_2^*/\varepsilon]} \|\pi(D_{t_0,z_0})\mathcal{U}_{1,II}(T)\|_{E^{s+2}_{\tau_2^*}},$$

we can conclude that

$$\|\varepsilon\pi(D_{t_0,z_0})\underline{\mathcal{U}}^\varepsilon_{1,II}\|_{B^{s+2}_{\tau_2^*}} \le \tau_2^*\|\mathcal{U}_{0,I}\|^2_{A^{s+2}_{\tau_2^*}}\|\mathcal{U}_{0,II}\|_{A^{s+2}_{\tau_2^*}}.$$

Thus, $\varepsilon\pi(D_{t_0,z_0})\underline{\mathcal{U}}^\varepsilon_{1,II}$ is in $B^{s+2}_{\tau_2^*}$ and is uniformly bounded with respect to $\varepsilon \in (0,1)$. The same thing can be said about $\underline{\mathcal{U}}^\varepsilon_{2,II}$, which proves that $\underline{\mathcal{U}^\varepsilon} \in B^{s+2}_{\tau_2^*}$, with uniformly bounded norm.

We recall that

$$\partial_\tau\underline{\mathcal{U}}^\varepsilon + \varepsilon^{-1}(A_1\partial_X + A_2\partial_Y + A_3\partial_Z)\underline{\mathcal{U}}^\varepsilon + \varepsilon^{-2}(\partial_{t_0} + A_3\partial_{z_0} + L_0)\underline{\mathcal{U}}^\varepsilon + f(\underline{\mathcal{U}}^\varepsilon) = \sum_{j=-1}^{7}\varepsilon^{j-1}\underline{\mathcal{R}}^\varepsilon_j,$$

where the profiles $\underline{\mathcal{R}}^\varepsilon_j$ are defined as $\underline{\mathcal{R}}^\varepsilon_j(\tau, X, Y, Z, t_0, z_0) = \mathcal{R}_j(\tau, \tau/\varepsilon, X, Y, Z, t_0, z_0)$ with $\mathcal{R}_j$ given in (1.8).

Thanks to the results of the previous section, we know that $\underline{\mathcal{R}}^\varepsilon_{-1} = \underline{\mathcal{R}}^\varepsilon_0 = 0$ and that $\underline{\mathcal{R}}^\varepsilon_1 = (1-\psi^\delta(D_{z_0}))F^S(\mathcal{U}_{0,I,1}, \overline{\mathcal{U}_{0,I,1}}, \mathcal{U}_{0,II})$. Therefore,

$$\|\underline{\mathcal{R}}^\varepsilon_1\|_{B^s_{\tau_2^*}} \le \|\mathcal{U}_{0,I}\|^2_{A^s_{\tau_2^*}}\|(1-\psi^\delta(D_{z_0}))\underline{\mathcal{U}}^\varepsilon_{0,II}\|_{B^s_{\tau_2^*}}.$$

The following lemma says that this term goes to zero as $\delta \to 0$ uniformly in $\varepsilon \in (0,1)$.

LEMMA 2.9. *Let $\sigma \ge s$ and $\mathcal{U}_{0,II} \in A^\sigma_{\tau_2^*}$ be as given by Proposition* 2.6; *let $\underline{\mathcal{U}}^\varepsilon_{0,II} \in B^\sigma_{\tau_2^*}$ be defined as $\underline{\mathcal{U}}^\varepsilon_{0,II}(\tau, X, Y, Z, t_0, z_0) = \mathcal{U}_{0,II}(\tau, \tau/\varepsilon, X, Y, Z, t_0, z_0)$.*

Then one has $(1-\psi^\delta(D_{z_0}))\underline{\mathcal{U}}^\varepsilon_{0,II} \to 0$ in $B^\sigma_{\tau_2^*}$ as $\delta \to 0$ uniformly in $\varepsilon \in (0,1)$.

Proof. Let $\lambda^\varepsilon := \mathcal{F}\underline{\mathcal{U}}^\varepsilon_{0,II} \in \mathcal{C}([0,\tau_2^*], \mathcal{BV}(\mathbb{R}^2_\xi, H^\sigma(\mathbb{R}^3_{X,Y,Z})^n))$ and $\lambda_0 := \mathcal{F}\mathbf{U}^0_{II}$. Thanks to (2.31), we have

$$\widehat{\lambda}^\varepsilon(\tau) = e^{-i\tau(\frac{\omega'(k)}{2k}(\eta_1^2+\eta_2^2)+\frac{\omega''(k)}{2}\eta_3^2)}e^{i\frac{\tau}{\varepsilon}\omega'(k)\eta_3}\widehat{\lambda}_0.$$

We also know by the Radon–Nikodým property that there exists an H^σ-valued integral function r_0 such that $\|r_0(\xi)\| = 1$ for $v(\lambda_0)$ for almost all $\xi = (\omega, k)$ and such that for all Borel sets $E \subset \mathbb{R}^2$,

$$\lambda_0(E) = \int_E r_0(\xi)v(\lambda_0)(d\xi),$$

where $v(\lambda_0)$ denotes the total variation measure associated to λ_0. Hence,

$$\widehat{\lambda}^\varepsilon(\tau)(E) = \int_E e^{-i\tau(\frac{\omega'(k)}{2k}(\eta_1^2+\eta_2^2)+\frac{\omega''(k)}{2}\eta_3^2)}e^{i\frac{\tau}{\varepsilon}\omega'(k)\eta_3}\widehat{r}_0(\xi)v(\lambda_0)(d\xi), \tag{2.33}$$

and also

$$(1-\psi^\delta(k))\widehat{\lambda}^\varepsilon(\tau)(E) = \int_E (1-\psi^\delta(k))e^{-i\tau(\frac{\omega'(k)}{2k}(\eta_1^2+\eta_2^2)+\frac{\omega''(k)}{2}\eta_3^2)}e^{i\frac{\tau}{\varepsilon}\omega'(k)\eta_3} \times\widehat{r}_0(\xi)v(\lambda_0)(d\xi).$$

Therefore

$$\|(1-\psi^\delta(D_{z_0}))\underline{\mathcal{U}}^\varepsilon_{0,II}\|_{B^\sigma_{\tau_2^*}} = \sup_{\tau\in[0,\tau_2^*]}|(1-\psi^\delta(k))\widehat{\lambda}^\varepsilon|_{\mathcal{B}V} \le \int_E (1-\psi^\delta(k))v(\lambda_0)(d\xi)$$

and hence tends to zero as $\delta \to 0$ uniformly in $\varepsilon \in (0,1)$ as a consequence of the dominated convergence theorem. ∎

We now turn to investigating the residual terms $\varepsilon^{j-1}\underline{\mathcal{R}}_j^\varepsilon$ for $j \geq 2$.

Using point (iii) of Lemma 2.7 and with the same techniques used above to prove that $\varepsilon\pi(D_{t_0,z_0})\underline{\mathcal{U}}_{1,II}^\varepsilon$ is uniformly bounded in $B^{s+2}_{\tau_2^*}$, one can prove that $\varepsilon\underline{\mathcal{R}}_j^\varepsilon$ is uniformly bounded in $B^s_{\tau_2^*}$ with respect to $\varepsilon \in (0,1)$. Therefore, if $j \geq 3$, then the residual terms $\varepsilon^{j-1}\underline{\mathcal{R}}_j^\varepsilon$ tend to 0 in $B^s_{\tau_2^*}$ when $\varepsilon \to 0$ (and with $\delta > 0$ being fixed).

The only remaining term to treat is therefore $\varepsilon\underline{\mathcal{R}}_2^\varepsilon$. In fact, only its continuous spectrum component needs care; so far, we know only that it is uniformly bounded in $B^s_{\tau_2^*}$. We recall that $\mathcal{R}_{2,II}$ is given by

$$\begin{aligned}\mathcal{R}_{2,II} = &-L_1(\partial)\mathcal{L}^{-1}(D_{t_0,z_0})A(\partial_{X,Y,Z})\mathcal{U}_{0,II} + L_1(\partial)\mathcal{L}^{-1}(D_{t_0,z_0})\psi^\delta(D_{z_0})f(\mathcal{U}_0)_{II}\\ &+iL_1(\partial)\mathcal{L}^{-1}(D_{t_0,z_0})A(\partial_{X,Y,Z})\pi(D_{t_0,z_0})\mathcal{U}_{1,II} + i\mathcal{L}^{-1}(D_{t_0,z_0})A(\partial_{X,Y,Z})\partial_\tau\mathcal{U}_{0,II}\\ &+\partial_\tau\pi(D_{t_0,z_0})\mathcal{U}_{1,II} + \big(f'(\mathcal{U}_0)(\mathcal{U}_1)\big)_{II}.\end{aligned}$$

In this expression, all the terms which do not involve $\pi(D_{t_0,z_0})\mathcal{U}_{1,II}$ are in $A^s_{\tau_2^*}$, and their contribution to $\underline{\mathcal{R}}_2^\varepsilon$ is therefore in $B^s_{\tau_2^*}$. Hence, the only possible problems come from the terms which involve $\pi(D_{t_0,z_0})\mathcal{U}_{1,II}$. We need the following lemma.

LEMMA 2.10. *Let $\sigma \geq s+4$ and $\pi(D_{t_0,z_0})\mathcal{U}_{1,II} \in E^\sigma_{\tau_2^*}$ be as given by Proposition* 2.6.

Then one has

$$\lim_{\varepsilon\to 0}\varepsilon\|\underline{\mathcal{U}}_{1,II}^\varepsilon\|_{B^\sigma_{\tau_2^*}} = 0 \quad \text{and} \quad \lim_{\varepsilon\to 0}\varepsilon\|\partial_\tau\underline{\mathcal{U}}_{1,II}^\varepsilon\|_{B^{\sigma-2}_{\tau_2^*}} = 0.$$

Proof. For all $\tau \in [0,\tau_2^*]$, let $\nu^\varepsilon(\tau) := \mathcal{F}\underline{\mathcal{U}}_{1,II}^\varepsilon(\tau)$. Thanks to (2.32), one then has

$$\widehat{\nu^\varepsilon}(\tau) = -\int_0^{\tau/\varepsilon} e^{-(\frac{\tau}{\varepsilon}-t)\omega'(k)\eta_3}\psi^\delta(k)\widehat{\mu}(\tau,t)dt, \tag{2.34}$$

where, for all (τ,T), $\mu(\tau,T)$ is defined as $\mu(\tau,T) := \mathcal{F}\big(F^S(\mathcal{U}_{0,I,1},\overline{\mathcal{U}_{0,I,1}},\mathcal{U}_{0,II})(\tau,T)\big)$. We also recall that $\widehat{\ }$ denotes the Fourier transform with respect to the variables (X,Y,Z).

As with (2.33), we can write $\lambda(\tau,T) := \mathcal{F}\mathcal{U}_{0,II}(\tau,T)$ in the form

$$\widehat{\lambda}(\tau,T) = \int_E e^{-i\tau(\frac{\omega'(k)}{2k}(\eta_1^2+\eta_2^2)+\frac{\omega''(k)}{2}\eta_3^2)}e^{iT\omega'(k)\eta_3}\widehat{r_0(\xi)}v(\lambda_0)(d\xi)$$

for all Borel sets $E \subset \mathbb{R}^2$.

Hence, $\widehat{\mu}(\tau,T)(E) \in H^\sigma(\mathbb{R}^3)$ is given, for all Borel sets $E \subset \mathbb{R}^2$ and $\eta \in \mathbb{R}^3$, by

$$\begin{aligned}\widehat{\mu}(\tau,T)(E)(\eta) = \int_E\int_{\mathbb{R}^3\times\mathbb{R}^3} F^S\Big(&\widehat{\mathcal{U}_{0,I,1}}(\eta-\eta'),\widehat{\overline{\mathcal{U}_{0,I,1}}}(\eta''-\eta'),\\ &e^{-i\tau(\frac{\omega'(k)}{2k}(\eta_1'^2+\eta_2'^2)+\frac{\omega''(k)}{2}\eta_3'^2)}e^{iT\omega'(k)\eta_3'}\widehat{r_0(\xi)}(\eta')\Big)\,d\eta'd\eta''v(\lambda_0)(d\xi).\end{aligned}$$

Combining this equation with (2.34) then yields

$$\begin{aligned}|\widehat{\nu^\varepsilon}(\tau)|_{\mathcal{BV}} \leq &\int_{\mathbb{R}^2}\|\int_{\mathbb{R}^3\times\mathbb{R}^3}\int_0^{\tau/\varepsilon} F^S\Big(\Big(\widehat{\mathcal{U}_{0,I,1}}(\cdot-\eta'),\widehat{\overline{\mathcal{U}_{0,I,1}}}(\eta''-\eta'),\\ &e^{-i\tau(\frac{\omega'(k)}{2k}(\eta_1'^2+\eta_2'^2)+\frac{\omega''(k)}{2}\eta_3'^2)}e^{it\omega'(k)\eta_3'}\widehat{r_0(\xi)}(\eta')\Big)\,dtd\eta'd\eta''\|_{\mathcal{F}(H^\sigma)}v(\lambda_0)(d\xi)\\ := &\int_{\mathbb{R}^2} G^\varepsilon(\xi)v(\lambda_0)(d\xi).\end{aligned}$$

The family $\varepsilon G^\varepsilon(\xi)$ can be bounded by a constant (and constants are $v(\lambda_0)$-integrable); thanks to Lemma 6 of [11], we also know that $\varepsilon G^\varepsilon(\xi) \to 0$ as $\varepsilon \to 0$, provided that $\omega'(k) \neq \omega'(k_l)$. Since this equality occurs only when $k = k_l$, and since $v(\lambda_0)(\{k_l\}) = 0$, we thus have $\varepsilon G^\varepsilon(\xi) \to 0$ $v(\lambda_0)$ almost everywhere, so that the first part of the lemma follows from the dominated convergence theorem.

The second part of the lemma can be established with the same techniques. ∎

From the above lemma, it is clear that $\varepsilon \underline{\mathcal{R}}_2^\varepsilon = o(1)$, which achieves the last step of the proof of the proposition. ∎

We know that u^ε almost solves (1.1), but we have not yet proved that the difference $\mathbf{u}^\varepsilon - u^\varepsilon$ remains small. This is what the following theorem shows.

THEOREM 2.11. *Suppose the characteristic variety $\mathcal{C}_\mathcal{L}$ is as in Assumptions* 2.1 *and* 2.3.

Let $\mathbf{U}^0 = \mathbf{U}_I^0 + \mathbf{U}_{II}^0 \in A_0^{s+4}$ such that $\mathbf{U}_I^0 = \mathbf{U}_{I,1}^0 e^{i\theta} + c.c.$ and suppose, moreover, that

$$\pi(\omega_l, -k_l)\mathbf{U}_{I,1}^0 = \mathbf{U}_{I,1}^0 \quad \textit{and} \quad \pi(D_{t_0,z_0})\mathbf{U}_{II}^0 = \mathbf{U}_{II}^0,$$

and that Sp $\mathbf{U}_{II}^0 \subset \{(\omega,k), |k| > \delta\}$ *for a given $\delta > 0$.*

Then, for $0 < \tau^* \leq \inf(\tau_1^*, \tau_2^*)$, the following holds:

(i) *The profile $\mathcal{U}_0 = \mathcal{U}_{0,I} + \mathcal{U}_{0,II}$ given by Propostion* 2.6 *satisfies Assumption* 2.2, *and the associated profile $\underline{\mathcal{U}}_0^\varepsilon \in B_{\tau^*}^s$ approximates the singular equation* (1.5) *in the sense that for all $\mu > 0$ there exists a $\delta(\mu)$ such that if $0 < \delta < \delta(\mu)$, then*

$$\|\mathbf{U}_I^\varepsilon - \underline{\mathcal{U}}_{0,I}^\varepsilon\|_{B_{\tau^*}^s} = O(\varepsilon) \quad \textit{and} \quad \limsup_{\varepsilon\to 0} \|\mathbf{U}_{II}^\varepsilon - \underline{\mathcal{U}}_{0,II}^\varepsilon\|_{B_{\tau^*}^s} < \mu/3,$$

where we have decomposed the profile $\mathbf{U}^\varepsilon$ of the exact solution $\mathbf{u}^\varepsilon$, given by Theorem 1.4, *into $\mathbf{U}^\varepsilon = \mathbf{U}_I^\varepsilon + \mathbf{U}_{II}^\varepsilon$.*

(ii) *We also have stability of the approximate solution defined with $\mathcal{U}_0$,*

$$\|\mathbf{u}_I^\varepsilon - u_{0,I}^\varepsilon\| = O(\varepsilon^{3/2}) \quad \textit{and} \quad \limsup_{\varepsilon\to 0} \frac{1}{\sqrt{\varepsilon}} \|\mathbf{u}_{II}^\varepsilon - u_{0,II}^\varepsilon\| < \mu/3,$$

where the norm can be taken either in $\mathcal{C}([0, \frac{\tau^*}{\varepsilon}] \times \mathbb{R}^3)^n$ or in $\mathcal{C}([0, \frac{\tau^*}{\varepsilon}], L^2(\mathbb{R}^3)^n)$.

Notation. We have used in this theorem the notation

$$\mathbf{u}_I^\varepsilon = \sqrt{\varepsilon}\mathbf{U}_I^\varepsilon(\varepsilon T, X, Y, Z, T/\varepsilon, Z/\varepsilon), \quad u_{0,I}^\varepsilon = \sqrt{\varepsilon}\underline{\mathcal{U}}_{0,I}^\varepsilon(\varepsilon T, X, Y, Z, T/\varepsilon, Z/\varepsilon),$$

and similarly

$$\mathbf{u}_{II}^\varepsilon = \sqrt{\varepsilon}\mathbf{U}_{II}^\varepsilon(\varepsilon T, X, Y, Z, T/\varepsilon, Z/\varepsilon), \quad u_{0,II}^\varepsilon = \sqrt{\varepsilon}\underline{\mathcal{U}}_{0,II}^\varepsilon(\varepsilon T, X, Y, Z, T/\varepsilon, Z/\varepsilon).$$

Proof. (i) Since $\underline{\mathcal{R}}_{-1}^\varepsilon = \underline{\mathcal{R}}_0^\varepsilon = 0$, the error profile $\mathcal{W}^\varepsilon = \mathbf{U}^\varepsilon - \underline{\mathcal{U}}^\varepsilon$ satisfies

$$\begin{aligned}\partial_\tau \mathcal{W}^\varepsilon + \varepsilon^{-1}(A_1\partial_X + A_2\partial_Y + A_3\partial_Z)\mathcal{W}^\varepsilon + \varepsilon^{-2}(\partial_{t_0} + A_3\partial_{z_0} + L_0)\mathcal{W}^\varepsilon \\ = f(\underline{\mathcal{U}}^\varepsilon) - f(\mathbf{U}^\varepsilon) + \underline{\mathcal{R}}_1^\varepsilon + \sum_{j=2}^7 \varepsilon^{j-1}\underline{\mathcal{R}}_j^\varepsilon.\end{aligned}$$

Thanks to Taylor's theorem, there exists a regular function G such that $f(\underline{\mathcal{U}}^\varepsilon) - f(\mathbf{U}^\varepsilon) = G(\underline{\mathcal{U}}^\varepsilon, \mathbf{U}^\varepsilon)\mathcal{W}^\varepsilon$. Therefore, the profile $\mathcal{W}^\varepsilon$ satisfies

$$\begin{aligned}\partial_\tau \mathcal{W}^\varepsilon + \varepsilon^{-1}(A_1\partial_X + A_2\partial_Y + A_3\partial_Z)\mathcal{W}^\varepsilon + \varepsilon^{-2}(\partial_{t_0} + A_3\partial_{z_0} + L_0)\mathcal{W}^\varepsilon \\ - G(\underline{\mathcal{U}}^\varepsilon, \mathbf{U}^\varepsilon)\mathcal{W}^\varepsilon = \underline{\mathcal{R}}_1^\varepsilon + \sum_{j=2}^7 \varepsilon^{j-1}\underline{\mathcal{R}}_j^\varepsilon.\end{aligned} \tag{2.35}$$

Let $R > 0$ such that $\|\mathbf{U}^0\|_{A^{s+4}_{\tau^*}} \leq R$. We recall that τ_1^* and τ_2^* are chosen (in Theorem 1.4 and Proposition 2.6, respectively) in such a way that $\|\mathbf{U}^\varepsilon\|_{A^s_{\tau_1^*}} \leq 2R$ and $\|\mathcal{U}_0\|_{A^{s+4}_{\tau_2^*}} \leq 2R$. Since $\tau^* \leq \inf(\tau_1^*, \tau_2^*)$, we can replace $\tau_{1,2}^*$ by τ^* in these inequalities. Hence, we can deduce from Lemma 2.7 that

$$\|\underline{\mathcal{U}}^\varepsilon\|_{B^s_{\tau^*}} \leq CR\left(R^2 + \varepsilon\left(1 + \frac{1+R^2}{\delta}\right) + \varepsilon^2(1+R^2)\left(\frac{1}{\delta} + \frac{1}{\delta^2}\right)\right).$$

Thus

$$\|G(\underline{\mathcal{U}}^\varepsilon, \mathbf{U}^\varepsilon)\|_{B^s_{\tau^*}} \leq h\left(R\left(R^2 + \varepsilon\left(1 + \frac{1+R^2}{\delta}\right) + \varepsilon^2(1+R^2)\left(\frac{1}{\delta} + \frac{1}{\delta^2}\right)\right), R\right),$$

where $h(\cdot,\cdot)$ is a smooth positive function which does not depend on R, δ, nor ε.

By a Gronwall-type argument, we can therefore deduce from (2.35) that

$$\|\mathcal{W}^\varepsilon\|_{B^s_{\tau^*}} \leq \tau^* \left(\|\underline{\mathcal{R}}_1^\varepsilon\|_{B^s_{\tau^*}} + \sum_{j=2}^{7} \varepsilon^{j-1}\|\underline{\mathcal{R}}_j^\varepsilon\|_{B^s_{\tau^*}}\right) e^{h\left(R(R^2+\varepsilon(1+\frac{1+R^2}{\delta})+\varepsilon^2(1+R^2)(\frac{1}{\delta}+\frac{1}{\delta^2})),R\right)\tau^*}. \tag{2.36}$$

It is now easy to deduce from (2.36), Lemma 2.7(iii), and Proposition 2.8 that $\underline{\mathcal{U}}^\varepsilon = \underline{\mathcal{U}}_0^\varepsilon + \varepsilon\underline{\mathcal{U}}_1^\varepsilon + \varepsilon^2\underline{\mathcal{U}}_2^\varepsilon$ satisfies the asymptotic properties of point (i) of the theorem. This point will be proved if we can replace $\underline{\mathcal{U}}^\varepsilon$ by $\underline{\mathcal{U}}_0^\varepsilon$. This is obviously the case since, as a consequence of Lemmas 2.7 and 2.10, $\varepsilon\underline{\mathcal{U}}_1^\varepsilon + \varepsilon^2\underline{\mathcal{U}}_2^\varepsilon$ goes to 0 in $B^s_{\tau^*}$ as $\varepsilon \to 0$.

Taking the discrete spectrum component of (2.35) yields the usual equations of diffractive optics (in particular, $\mathcal{R}_{1,I} = 0$) so that the techniques of [8], [10], and [11] can be used to obtain a better estimate $O(\varepsilon)$ of the error term.

(ii) This point is a direct consequence of (i) and of the embedding results of Proposition 1.3. ∎

2.5. Stability in the general case. In this section, we consider the general case, i.e., we allow low frequencies. Therefore, we consider initial conditions with profile $\mathbf{U}^0 = \mathbf{U}_I^0 + \mathbf{U}_{II}^0 \in A_0^{s+4}$ without making any assumption on the spectrum of $\mathbf{U}_{II}^0$. Alterman's methods [1] are used to relax this assumption. We first introduce the following notation.

Notation. We denote by $\mathbf{U}_{II}^{0,\delta}$ and $\mathbf{U}^{0,\delta}$ the "filtered" profiles

$$\mathbf{U}_{II}^{0,\delta} = \psi^\delta(D_{z_0})\mathbf{U}_{II}^0 \quad \text{and} \quad \mathbf{U}^{0,\delta} = \mathbf{U}_I^0 + \mathbf{U}_{II}^{0,\delta}, \tag{2.37}$$

where $0 < \delta < 1$.

The exact solution of (1.1) with initial condition $\sqrt{\varepsilon}\mathbf{U}^{0,\delta}(X,Y,Z,0,Z/\varepsilon)$ determined by Theorem 1.4 is denoted by $\mathbf{u}^{\varepsilon,\delta}$ and its associated profile by $\mathbf{U}^{\varepsilon,\delta}$, so that one has $\mathbf{u}^{\varepsilon,\delta}(T,X,Y,Z) = \sqrt{\varepsilon}\mathbf{U}^{\varepsilon,\delta}(\varepsilon T, X, Y, Z, T/\varepsilon, Z/\varepsilon)$.

The dominated convergence theorem shows that $\mathbf{U}_{II}^{0,\delta} \to \mathbf{U}_{II}^0$ in A_0^s. We also have convergence of the exact solutions of (1.1) associated to these initial conditions, as the following proposition shows.

PROPOSITION 2.12. *Let $\mathbf{U}^0 = \mathbf{U}_I^0 + \mathbf{U}_{II}^0 \in A_0^s$ and $\mathbf{U}^{0,\delta} = \mathbf{U}_I^0 + \psi^\delta(D_{z_0})\mathbf{U}_{II}^0$.*

There exists $\tau_1^* > 0$, independent of ε and δ, such that the exact solutions $\mathbf{U}^\varepsilon$ and $\mathbf{U}^{\varepsilon,\delta}$ of the singular equation (1.5) *with initial conditions $\mathbf{U}^0$ and $\mathbf{U}^{0,\delta}$, respectively, exist in $B^s_{\tau_1^*}$. Moreover, one has*

$$\mathbf{U}^\varepsilon - \mathbf{U}^{\varepsilon,\delta} \to 0 \quad \text{in } B^s_{\tau_1^*} \quad \text{as } \delta \to 0$$

uniformly in $\varepsilon \in (0,1)$.

Proof. It is easy to see that $\|\mathbf{U}^{0,\delta}\|_{B^s_{\tau_1^*}} \leq \|\mathbf{U}^0\|_{B^s_{\tau_1^*}}$. Hence, if R is such that $\|\mathbf{U}^0\|_{B^s_{\tau_1^*}} \leq R$, one also has $\|\mathbf{U}^{0,\delta}\|_{B^s_{\tau_1^*}} \leq R$. Therefore, Theorem 1.4 implies that the profile $\mathbf{U}^{\varepsilon,\delta}$ also exists on the existence interval $[0,\tau_1^*]$ of $\mathbf{U}^\varepsilon$, since this interval depends only on R.

Moreover, on $[0,\tau_1^*]$, the difference $\mathbf{W}^\delta = \mathbf{U}^\varepsilon - \mathbf{U}^{\varepsilon,\delta}$ satisfies

$$\begin{aligned}\partial_\tau \mathbf{W}^\delta + \varepsilon^{-1}(A_1\partial_X + A_2\partial_Y + A_3\partial_Z)\mathbf{W}^\delta + \varepsilon^{-2}(\partial_{t_0} + A_3\partial_{z_0} + L_0)\mathbf{W}^\delta \\ = G(\mathbf{U}^{\varepsilon,\delta}, \mathbf{U}^\varepsilon)\mathbf{W}^\delta,\end{aligned}$$

where, as for (2.35), G is a regular function satisfying $G(\mathbf{U}^{\varepsilon,\delta}, \mathbf{U}^\varepsilon)\mathbf{W}^\delta = f(\mathbf{U}^{\varepsilon,\delta}) - f(\mathbf{U}^\varepsilon)$.

As in the proof of Theorem 2.11, we obtain

$$\|G(\mathbf{U}^{\varepsilon,\delta}, \mathbf{U}^\varepsilon)\|_{B^s_{\tau_1^*}} \leq h(R,R),$$

where $h(\cdot,\cdot)$ is a smooth positive function independent of δ and ε.

A Gronwall-type argument then yields

$$\|\mathbf{W}^\delta\|_{B^s_{\tau_1^*}} \leq \|\mathbf{W}^{0,\delta}\|_{A^s_0} e^{h(R,R)\tau_1^*},$$

so that the desired result is now a consequence of the dominated convergence theorem. ∎

We now study the convergence of the approximate solutions. If we take $\mathbf{U}^{0,\delta}$ as the initial condition, all the results of sections 2.1–2.4 remain valid. In particular, we can construct an approximate profile $\underline{\mathcal{U}}^{\varepsilon,\delta} = \underline{\mathcal{U}}_0^{\varepsilon,\delta} + \varepsilon\underline{\mathcal{U}}_1^{\varepsilon,\delta} + \varepsilon^2\underline{\mathcal{U}}_2^{\varepsilon,\delta}$ of $\mathbf{U}^{\varepsilon,\delta}$. The leading term $\underline{\mathcal{U}}_0^{\varepsilon,\delta}$ satisfies

$$\underline{\mathcal{U}}_0^{\varepsilon,\delta}(\tau, X, Y, Z, t_0, z_0) = \mathcal{U}_0^\delta(\tau, \tau/\varepsilon, X, Y, Z, t_0, z_0),$$

with $\mathcal{U}_0^\delta = \mathcal{U}_{0,I} + \mathcal{U}_{0,II}^\delta$. The discrete spectrum component $\mathcal{U}_{0,I}$ (which does not depend on δ) is given as before by system (2.27), while $\mathcal{U}_{0,II}^\delta = \pi(D_{t_0,z_0})\mathcal{U}_{0,II}^\delta$ is found solving

$$\left\{\begin{array}{l}(\partial_T - \omega'(D_{z_0})\partial_Z)\mathcal{U}_{0,II}^\delta = 0,\\ \partial_\tau \mathcal{U}_{0,II}^\delta + i\frac{\omega'(D_{z_0})}{2D_{z_0}}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,II}^\delta + i\frac{\omega''(D_{z_0})}{2}\partial_Z^2\mathcal{U}_{0,II}^\delta = 0,\\ \mathcal{U}_{0,II}^\delta|_{\tau=T=0} = \mathbf{U}_{II}^{0,\delta}.\end{array}\right. \tag{2.38}$$

The following proposition shows that when $\delta \to 0$, the profile $\mathcal{U}_{0,II}^\delta$ tends to the profile $\mathcal{U}_{0,II}$ obtained formally by taking $\delta = 0$ in (2.38).

PROPOSITION 2.13. *Let $\sigma \geq s$ such that $\mathbf{U}_{II}^0 \in A_0^\sigma$. Suppose, moreover, that $\pi(D_{t_0,z_0})\mathbf{U}_{II}^0 = \mathbf{U}_{II}^0$.*

Then there exists $\tau_2^* > 0$ such that the solution $\mathcal{U}_{0,II}^\delta$ of (2.38) exists in $A^\sigma_{\tau_2^*}$ for all $0 < \delta < 1$ and such that the limit system

$$\left\{\begin{array}{l}(\partial_T - \omega'(D_{z_0})\partial_Z)\mathcal{U}_{0,II} = 0,\\ \partial_\tau\partial_{z_0}\mathcal{U}_{0,II} - \frac{\omega'(D_{z_0})}{2}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,II} - \frac{D_{z_0}\omega''(D_{z_0})}{2}\partial_Z^2\mathcal{U}_{0,II} = 0,\\ \mathcal{U}_{0,II}|_{\tau=T=0} = \mathbf{U}_{II}^0\end{array}\right. \tag{2.39}$$

admits a unique solution in $A^\sigma_{\tau_2^*}$.

Moreover, $\mathcal{U}^\delta_{0,II} \to \mathcal{U}_{0,II}$ *in* $A^\sigma_{\tau_2^*}$ *as* $\delta \to 0$.

Proof. The results of the proposition are easily obtained from the explicit expression of $\mathcal{U}^\delta_{0,II}$ given by (2.31) and from the dominated convergence theorem. ∎

Remark 2.10. (i) Since $\omega'(k)$ is an even function and $\omega''(k)$ an odd function of (ω, k), the Fourier multipliers $\omega'(D_{z_0})$ and $D_{z_0}\omega''(D_{z_0})$ transform real functions of the variable (t_0, z_0) into real functions.

(ii) System (2.39) is formally obtained by differentiating (2.38) and letting $\delta \to 0$. In (2.39), there is no D_{z_0} inverse (which is not a Fourier multiplier), and therefore this system can be solved explicitly in the Fourier domain.

We are now ready to state our main theorem.

THEOREM 2.14. *Suppose the characteristic variety* $\mathcal{C}_\mathcal{L}$ *is as in Assumptions* 2.1 *and* 2.3.

Let $\mathbf{U}^0 = \mathbf{U}^0_I + \mathbf{U}^0_{II} \in A^{s+4}_0$ *such that* $\mathbf{U}^0_I = \mathbf{U}^0_{I,1} e^{i\theta} + c.c.$ *and suppose, moreover, that*

$$\pi(\omega_l, -k_l)\mathbf{U}^0_{I,1} = \mathbf{U}^0_{I,1} \quad \textit{and} \quad \pi(D_{t_0,z_0})\mathbf{U}^0_{II} = \mathbf{U}^0_{II}.$$

Then for $\tau_3^* = \min\{\tau_1^*, \tau_2^*\}$ *we have*

(i) *the exact solution* $\mathbf{u}^\varepsilon$ *of* (1.1) *exists on* $[0, \tau_3^*/\varepsilon]$ *and can be written* $\mathbf{u}^\varepsilon(T, X, Y, Z) = \sqrt{\varepsilon}\mathbf{U}^\varepsilon(\varepsilon T, X, Y, Z, T/\varepsilon, Z/\varepsilon)$, *with* $\mathbf{U}^\varepsilon = \mathbf{U}^\varepsilon_I + \mathbf{U}^\varepsilon_{II} \in B^{s+4}_{\tau_3^*}$;

(ii) $\mathcal{U}_{0,I,1}$ *is defined in* $\mathcal{C}_b([0, \tau_3^*] \times \mathbb{R}_T, H^{s+4}(\mathbb{R}^3)^n)$ *as the unique solution of* (2.27);

(iii) $\mathcal{U}_{0,II}$ *is defined in* $A^{s+4}_{\tau_3^*}$ *as the unique solution of* (2.39);

(iv) *the profile* $\underline{\mathcal{U}}^\varepsilon_0 \in B^s_{\tau_3^*}$ *associated to* $\mathcal{U}_0 = \mathcal{U}_{0,I} + \mathcal{U}_{0,II} \in A^s_{\tau_3^*}$, *with* $\mathcal{U}_{0,I} = \mathcal{U}_{0,I,1}e^{i\theta} + c.c.$, *approximates the singular equation* (1.5) *in the sense that*

$$\|\mathbf{U}^\varepsilon_I - \underline{\mathcal{U}}^\varepsilon_{0,I}\|_{B^s_{\tau_3^*}} = O(\varepsilon) \quad \textit{and} \quad \|\mathbf{U}^\varepsilon_{II} - \underline{\mathcal{U}}^\varepsilon_{0,II}\|_{B^s_{\tau_3^*}} = o(1) \quad \textit{as } \varepsilon \to 0;$$

(v) *we also have stability of the approximate solution* u^ε_0 *defined with* $\mathcal{U}_0$,

$$\|\mathbf{u}^\varepsilon_I - u^\varepsilon_{0,I}\| = O(\varepsilon^{3/2}) \quad \textit{and} \quad \|\mathbf{u}^\varepsilon_{II} - u^\varepsilon_{0,II}\| = o(\sqrt{\varepsilon}) \quad \textit{as } \varepsilon \to 0,$$

where the norm can be taken either in $\mathcal{C}([0, \frac{\tau_3^*}{\varepsilon}] \times \mathbb{R}^3)^n$ *or in* $\mathcal{C}([0, \frac{\tau_3^*}{\varepsilon}], L^2(\mathbb{R}^3)^n)$.

Notation. We have used the same notation $\mathbf{u}^\varepsilon_I$, $\mathbf{u}^\varepsilon_{II}$, $u^\varepsilon_{0,I}$, and $u^\varepsilon_{0,II}$ as in Theorem 2.11.

Proof. (i)–(iii) The first three points have been proved in Theorem 1.4, Proposition 2.6, and Proposition 2.13.

(iv) Convergence of the discrete spectrum components is exactly the same as in Theorem 2.11 since the assumption of the absence of low frequencies only affects the continuous spectrum components.

We now want to prove that $\mathbf{U}^\varepsilon_{II} \to \underline{\mathcal{U}}^\varepsilon_{0,II}$ in $B^s_{\tau_3^*}$, i.e., for all $\mu > 0$, there exists $\varepsilon_0 > 0$ such that for $0 < \varepsilon < \varepsilon_0$, $\|\mathbf{U}^\varepsilon_{II} - \underline{\mathcal{U}}^\varepsilon_{0,II}\|_{B^s_{\tau_3^*}} < \mu$.

Now, write

$$\|\mathbf{U}^\varepsilon_{II} - \underline{\mathcal{U}}^\varepsilon_{0,II}\|_{B^s_{\tau_3^*}} \le \|\mathbf{U}^\varepsilon_{II} - \mathbf{U}^{\varepsilon,\delta}_{II}\|_{B^s_{\tau_3^*}} + \|\mathbf{U}^{\varepsilon,\delta}_{II} - \underline{\mathcal{U}}^{\varepsilon,\delta}_{0,II}\|_{B^s_{\tau_3^*}} + \|\underline{\mathcal{U}}^{\varepsilon,\delta}_{0,II} - \underline{\mathcal{U}}^\varepsilon_{0,II}\|_{B^s_{\tau_3^*}},$$

where $\delta > 0$ and $\mathbf{U}^{\varepsilon,\delta}_{II}$ and $\underline{\mathcal{U}}^{\varepsilon,\delta}_{0,II}$ are defined as usual.

Thanks to Propositions 2.12–2.13, we know that for $\delta \le \delta_0$ small enough and for all $\varepsilon \in (0, 1)$, one has

$$\|\mathbf{U}^\varepsilon_{II} - \mathbf{U}^{\varepsilon,\delta}_{II}\|_{B^s_{\tau_3^*}} < \mu/3 \quad \text{and} \quad \|\underline{\mathcal{U}}^{\varepsilon,\delta}_{0,II} - \underline{\mathcal{U}}^\varepsilon_{0,II}\|_{B^s_{\tau_3^*}} \le \|\mathcal{U}^\delta_{0,II} - \mathcal{U}_{0,II}\|_{A^s_{\tau_3^*}} < \mu/3.$$

Moreover, the profile $\underline{\mathcal{U}}_{0,II}^{\varepsilon,\delta}$ satisfies all the assumptions required to apply Theorem 2.11. Therefore, taking $0 < \delta < \inf\{\delta_0, \delta(\mu)\}$, we know that for ε small enough,

$$\|\mathbf{U}_{II}^{\varepsilon,\delta} - \underline{\mathcal{U}}_{0,II}^{\varepsilon,\delta}\|_{B^s_{\tau_3^*}} < \mu/3.$$

The above three inequalities thus yield

$$\|\mathbf{U}_{II}^{\varepsilon} - \underline{\mathcal{U}}_{0,II}^{\varepsilon}\|_{B^s_{\tau_3^*}} < \mu,$$

which proves the result.

(v) This point is a direct consequence of point (iv) and of the embedding properties of Proposition 1.3. ∎

3. Nondispersive case. In section 2, we have considered dispersive systems. If most of the physical applications fall into this class, nondispersive systems are also physically relevant. Ultrashort pulses, for instance, are often modeled with such a framework. There is also a mathematical reason why we study nondispersive problems in this section: We have seen in the previous section that interactions between oscillations with a purely continuous spectrum are not possible. The proof of this result relies strongly on the dispersive properties of the characteristic variety. We show in this section that when these properties do not hold, nonlinearities can be observed on the continuous spectrum components.

As already mentioned, this nondispersive framework has already been investigated by Alterman and Rauch in [1], [2], [3], and [15]. Of course, our results coincide with theirs, but our method is completely different, and the nondispersive case appears to be a particular case of the general framework presented in this paper and does not require an ad hoc analysis.

The systems we consider here are in the form

$$\left\{ \begin{array}{l} L(\partial)\mathbf{u}^\varepsilon + f(\mathbf{u}^\varepsilon) = 0, \mathbf{u}^\varepsilon|_{T=0}(X,Y,Z) = \mathbf{u}_\varepsilon^0(X,Y,Z), \end{array} \right.$$

with $L(\partial) = A_0\partial_T + A_1\partial_X + A_2\partial_Y + A_3\partial_Z$. We thus consider problems of type (1.1) with $L_0 = 0$. As we have seen in Remark 1.2, we can suppose that $A_0 = Id$.

The symbol $\mathcal{L}(\omega, k)$ then reads $\mathcal{L}(\omega, k) = \omega Id + A_3 k$ and is therefore homogeneous of degree one in (ω, k) so that Assumption 2.1 is never realized. Nevertheless without any additional hypothesis on $L(\partial)$, we know some properties on the characteristic variety $\mathcal{C}_\mathcal{L}$. Since $\mathcal{L}(\omega, k)$ is homogeneous of degree one, $\mathcal{C}_\mathcal{L}$ is a union of lines which all go through the origin. Moreover, if (ω, k) and (ω', k') are on the same line of $\mathcal{C}_\mathcal{L}$, then one has $\pi(\omega, k) = \pi(\omega', k')$. From this point onward, we use the following notation.

Notation. We denote by $\mathcal{D}_1, \ldots, \mathcal{D}_N$ the lines such that $\mathcal{C}_\mathcal{L} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_N$. We denote by $-v_j$ the slope of line $\mathcal{D}_j$. If $(\omega, k) \in \mathcal{D}_j$, write $\pi_j := \pi(\omega, k)$ and $\mathcal{L}_j^{-1} := \mathcal{L}^{-1}(\omega, k)$. Up to a renumbering, we can also suppose that $(\omega_l, -k_l) \in \mathcal{D}_1$.

As the study of the nondispersive case does not raise many other difficulties than in the dispersive case, most of the following results are given without proof.

3.1. The profile equations. As in section 2, an approximate solution is sought in the form

$$u^\varepsilon(T, X, Y, Z) = \sqrt{\varepsilon}\mathcal{U}^\varepsilon\left(\varepsilon T, T, X, Y, Z, \frac{T}{\varepsilon}, \frac{Z}{\varepsilon}\right),$$

where the profile $\mathcal{U}^\varepsilon$ is written

$$\mathcal{U}^\varepsilon = \mathcal{U}_0 + \varepsilon \mathcal{U}_1 + \varepsilon^2 \mathcal{U}_2,$$

with $\mathcal{U}_0, \mathcal{U}_1, \mathcal{U}_2 \in E^s_{\tau^*}$.

Thanks to Proposition 1.10, these profiles are decomposed into a component with a discrete spectrum and a component with a purely continuous one,

$$\mathcal{U}_j = \mathcal{U}_{j,I} + \mathcal{U}_{j,II}, \qquad j = 0, 1, 2.$$

We recall that the profiles which are labeled I always have a discrete spectrum and the profiles which are labeled II always have a purely continuous one.

The analysis of the discrete spectrum components slightly differs from the analysis performed for the dispersive case. Indeed, the superior harmonics created by the nonlinearity are noncharacteristic in the dispersive case and thus do not play any important role because they are not propagated; conversely, in the nondispersive case, the superior harmonics are characteristic, and so we must seek $\mathcal{U}_{0,I}, \mathcal{U}_{1,I}$, and $\mathcal{U}_{2,I}$ as periodic functions. Since the nonlinearity is odd, only odd harmonics are created by the nonlinearity, if no even harmonic is present initially. Therefore, we look for profiles of the form

$$\mathcal{U}_{l,I}(\tau, T, X, Y, Z, \theta) = \sum_{j \in Z} \mathcal{U}_{l,I,2j+1}(\tau, T, X, Y, Z) e^{i(2j+1)\theta}, \qquad l = 0, 1, 2.$$

In order for these profiles to be in $E^s_{\tau^*}$, one must have normal convergence of the harmonics, and that is why we introduce the following spaces.

DEFINITION 3.1. *We denote by D^s_0 (resp., $D^s_{\tau^*}$) the set of the sequences of profiles $(\mathcal{V}_{2j+1})_{j \in Z}$, with $\mathcal{V}_{2j+1} \in H^s(\mathbb{R}^3)^n$ (resp., $\mathcal{C}_b([0, \tau^*] \times \mathbb{R}_T, H^s(\mathbb{R}^3)^n)$) and such that*

$$\sum_{j \in Z} ||\mathcal{V}_{2j+1}|| < \infty,$$

where $||.||$ represents the norm of $H^s(\mathbb{R}^3)^n$ (resp., $\mathcal{C}_b([0, \tau^*] \times \mathbb{R}_T, H^s(\mathbb{R}^3)^n)$).

This finite positive number endows D^s_0 (resp., $D^s_{\tau^*}$) with a norm, denoted by $\|\cdot\|_{D^s_0}$ (resp., $\|\cdot\|_{D^s_{\tau^*}}$).

Annihilating $\mathcal{R}_{-1,I}$ yields as usual the polarization condition

$$\pi_1 \mathcal{U}_{0,I,2j+1} = \mathcal{U}_{0,I,2j+1} \quad \forall j \in Z. \tag{3.1}$$

As in section 2.1.2 and thanks to Lemma 1.7, the annihilation of $\mathcal{R}_{0,I}$ is equivalent to

$$\begin{cases} \pi_1 L_1(\partial) \pi_1 \mathcal{U}_{0,I,2j+1} = 0, \\ (Id - \pi_1) \mathcal{U}_{1,I,2j+1} = \dfrac{i}{2j+1} \mathcal{L}_1^{-1} A(\partial_{X,Y,Z}) \mathcal{U}_{0,I,2j+1}, \end{cases} \tag{3.2}$$

and as in the dispersive case, we can impose

$$\pi_1 \mathcal{U}_{1,I,2j+1} = 0 \quad \forall j \in Z. \tag{3.3}$$

When we annihilate $\mathcal{R}_{1,I}$ the need for periodic functions appears clearly. Indeed, in the dispersive case, all harmonics different from $\pm\theta$ are solved by elliptic inversion.

This kind of inversion is not possible in the nondispersive case because all harmonics are characteristic. Since all harmonics are odd, the nonlinearity $f(\mathcal{U}_0)_I$ can be written

$$f(\mathcal{U}_0)_I = \Lambda(\mathcal{U}_{0,I,.}) = \sum_{j\in Z} \Lambda_{2j+1}(\mathcal{U}_{0,I,.})e^{i(2j+1)\theta}, \tag{3.4}$$

where the notation $\mathcal{U}_{0,I,.}$ stands for the sequence $(\mathcal{U}_{0,I,2j+1})_{j\in Z}$. The annihilation of $\mathcal{R}_{1,I}$ is then equivalent to

$$\begin{aligned} i(2j+1)\mathcal{L}(\omega_l,-k_l)\mathcal{U}_{2,I,2j+1} + L_1(\partial)\mathcal{U}_{1,I,2j+1} + \partial_\tau \mathcal{U}_{0,I,2j+1} \\ + \Lambda_{2j+1}(\mathcal{U}_{0,I,.}) = 0 \end{aligned}$$

for all $j \in Z$. These equations are decomposed, thanks to Lemma 1.7 and (3.2)–(3.3), into

$$\begin{aligned} (Id-\pi_1)\mathcal{U}_{2,I,2j+1} = &-\frac{1}{(2j+1)^2}\mathcal{L}_1^{-1}L_1(\partial)\mathcal{L}_1^{-1}A(\partial_{X,Y,Z})\mathcal{U}_{0,I,2j+1} \\ &+ \frac{i}{2j+1}\mathcal{L}_1^{-1}\big(\partial_\tau \mathcal{U}_{0,I,2j+1} + \Lambda_{2j+1}(\mathcal{U}_{0,I,.})\big) \end{aligned}$$

and

$$\partial_\tau \mathcal{U}_{0,I,2j+1} + i\pi_1 A(\partial_{X,Y,Z})\mathcal{L}_1^{-1}A(\partial_{X,Y,Z})\pi_1\mathcal{U}_{0,I,2j+1} + \pi_1\Lambda_{2j+1}(\mathcal{U}_{0,I,.}) = 0.$$

Under Assumption 2.3, the same simplifications as in section 2.1.4 can be made using (3.1)–(3.3) and Proposition 2.4, so that $\mathcal{U}_{0,I,.}$ is found solving

$$\begin{cases} \pi_1\mathcal{U}_{0,I,.} = \mathcal{U}_{0,I,.}, \\ (\partial_T + v_1\partial_Z)\mathcal{U}_{0,I,.} = 0, \\ \partial_\tau \mathcal{U}_{0,I,.} + i\dfrac{v_1}{2k_l}(\partial_X^2+\partial_Y^2)\mathcal{U}_{0,I,.} + \pi_1\Lambda(\mathcal{U}_{0,I,.}) = 0 \end{cases} \tag{3.5}$$

in $D^s_{\tau^*}$.

If the nonlinearities are not studied in detail, the analysis of the components with a purely continuous spectrum is strictly the same as in the dispersive case. Provided that the continuous spectrum component $\mathcal{U}_{0,II}$ of $\mathcal{U}_0$ satisfies Assumption 2.2 (absence of low frequencies), we find as in section 2 that $\mathcal{U}_{0,II}$ must satisfy

$$\begin{cases} \pi(D_{t_0,z_0})\mathcal{U}_{0,II} = \mathcal{U}_{0,II}, \\ (\partial_T - \omega'(D_{z_0})\partial_Z)\mathcal{U}_{0,II} = 0, \\ \partial_\tau \mathcal{U}_{0,II} + i\dfrac{\omega'(D_{z_0})}{2D_{z_0}}(\partial_X^2+\partial_Y^2)\mathcal{U}_{0,II} + (\partial_T - \omega'(D_{z_0})\partial_Z)\pi(D_{t_0,z_0})\mathcal{U}_{1,II} \\ \qquad +\pi(D_{t_0,z_0})\psi^\delta(D_{z_0})[f(\mathcal{U}_0)]_{II} = 0, \end{cases} \tag{3.6}$$

where ψ^δ denotes the infrared cutoff introduced in Definition 2.2.

Yet, we can still simplify these equations in decomposing $\mathcal{U}_{0,II}$ in the form

$$\mathcal{U}_{0,II} = \mathcal{U}_{0,II,1} + \cdots + \mathcal{U}_{0,II,N}$$

such that the spectrum of $\mathcal{U}_{0,II,j}$ is included in $\mathcal{D}_j$ for all $j = 1,\dots,N$.

Hence, (3.6) read

$$\begin{cases} \pi_j \mathcal{U}_{0,II,j} = \mathcal{U}_{0,II,j}, \quad j = 1, \dots, N, \\ (\partial_T + v_j \partial_Z)\mathcal{U}_{0,II,j} = 0, \quad j = 1, \dots, N, \\ \partial_\tau \mathcal{U}_{0,II,j} - i\dfrac{v_j}{2D_{z_0}}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,II,j} + (\partial_T + v_j\partial_Z)\pi_j \mathcal{U}_{1,II,j} \\ \qquad + \pi_j \psi^\delta(D_{z_0})[f(\mathcal{U}_0)]_{II,j} = 0, \quad j = 1, \dots, N, \end{cases}$$

where we recall that $-v_j$ is the slope of line $\mathcal{D}_j$.

We now study the nonlinearity $\pi_j[f(\mathcal{U}_0)]_{II,j}$ which appears in the profile equations. With the same kind of argument as in the proof of Lemma 2.5, we can obtain the following lemma.

LEMMA 3.2. *Let $\mathcal{V}_{II,j} \in A^s_{\tau^*}$, $j = 1, \dots, N$, be N profiles with a purely continuous spectrum such that* Sp $\mathcal{V}_{II,j} \subset \mathcal{D}_j$. *Take also $a, b \in \mathbb{C}^n$. Then one has*

(i) $\pi_j F(\mathcal{V}_{II,k}, \mathcal{V}_{II,l}, \mathcal{V}_{II,m}) = 0$, *unless $j = k = l = m$;*
(ii) $\pi_j F(ae^{ik\theta}, \mathcal{V}_{II,l}, \mathcal{V}_{II,m}) = 0$ *for all $k \in Z$ unless $j = k = l = m = 1$;*
(iii) $\pi_j F(ae^{ik\theta}, be^{il\theta}, \mathcal{V}_{II,m}) = 0$ *for all $(k, l) \in Z^2$, unless $l + k = 0$ and $j = m$.*

Remark 3.1. The main difference between the dispersive and nondispersive cases is that we can have nonzero interactions between oscillations with a purely continuous spectrum in the nondispersive case. What the lemma says is that, in order to produce a nonzero interaction, these oscillations must have support on the same line as the characteristic variety. Therefore, the evolution equations of the modes $\mathcal{U}_{0,II,j}$ can be nonlinear.

Thanks to Lemma 3.2, the nonlinearity $\pi_j[f(\mathcal{U}_0)]_{II,j}$ may be written in the form

$$\begin{aligned} \pi_1[f(\mathcal{U}_0)]_{II,1} = \sum_{k\in Z} \pi_1 F^S(\mathcal{U}_{0,I,2k+1}, \overline{\mathcal{U}_{0,I,2k+1}}, \mathcal{U}_{0,II,1}) \\ + \pi_1 f'(\mathcal{U}_{0,II,1})(\mathcal{U}_{0,I}) + \pi_1 f(\mathcal{U}_{0,II,1}), \end{aligned}$$

and, when $j \geq 2$,

$$\pi_j[f(\mathcal{U}_0)]_{II,j} = \sum_{k\in Z} \pi_j F^S(\mathcal{U}_{0,I,2k+1}, \overline{\mathcal{U}_{0,I,2k+1}}, \mathcal{U}_{0,II,1}) + \pi_j f(\mathcal{U}_{0,II,j}).$$

3.2. Solving the profile equations. Inspired here again by [10] and [11], and using the expression of the nonlinearities given above, we decompose the equations on $\mathcal{U}_{0,II,j}$ as follows:

$$\begin{cases} \pi_1 \mathcal{U}_{0,II,1} = \mathcal{U}_{0,II,1}, \\ (\partial_T + v_1\partial_Z)\mathcal{U}_{0,II,1} = 0, \\ \partial_\tau \mathcal{U}_{0,II,1} - i\dfrac{v_1}{2D_{z_0}}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,II,1} \\ \qquad + \psi^\delta(D_{z_0})\pi_1 \displaystyle\sum_{k\in Z} F^S(\mathcal{U}_{0,I,2k+1}, \overline{\mathcal{U}_{0,I,2k+1}}, \mathcal{U}_{0,II,1}) \\ \qquad + \psi^\delta(D_{z_0})\pi_1 f'(\mathcal{U}_{0,II,1})(\mathcal{U}_{0,I}) + \psi^\delta(D_{z_0})\pi_1 f(\mathcal{U}_{0,II,1}) = 0 \\ (\partial_T + v_1\partial_Z)\mathcal{U}_{1,II,1} = 0 \end{cases} \tag{3.7}$$

and for $j \geq 2$,

$$
(3.8)\quad \begin{cases} \pi_j \mathcal{U}_{0,II,j} = \mathcal{U}_{0,II,j}, \\ (\partial_T + v_j \partial_Z)\mathcal{U}_{0,II,j} = 0, \\ \partial_\tau \mathcal{U}_{0,II,j} - i\dfrac{v_j}{2D_{z_0}}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,II,j} + \psi^\delta(D_{z_0})\pi_j f(\mathcal{U}_{0,II,j}) = 0 \\ (\partial_T + v_j \partial_Z)\pi_j \mathcal{U}_{1,II,j} = -\psi^\delta(D_{z_0})\pi_j \displaystyle\sum_{k\in Z} F^S(\mathcal{U}_{0,I,2k+1}, \overline{\mathcal{U}_{0,I,2k+1}}, \mathcal{U}_{0,II,1}). \end{cases}
$$

These profile equations can be solved. However, we will restrict ourselves to the case where $\mathcal{U}_{0,II} = \mathcal{U}_{0,II,1}$, i.e., where the spectrum of $\mathcal{U}_{0,II}$ is included in $\mathcal{D}_1$. The general case would be technically more difficult and is irrelevant for the physical examples considered in this paper.

PROPOSITION 3.3. *Let $\sigma \geq s$ and $\mathbb{R} > 0$ such that $\mathbf{U}^0 = \mathbf{U}^0_I + \mathbf{U}^0_{II,1} \in A^\sigma_0$ and $\|\mathbf{U}^0\|_{A^\sigma_0} \leq R$. Suppose, moreover, that $\mathbf{U}^0_I = \sum_{j\in Z} \mathbf{U}^0_{I,2j+1} e^{i(2j+1)\theta}$ with $\mathbf{U}^0_{I,\cdot} \in D^\sigma_0$. Assume finally that*

$$
\pi_1 \mathbf{U}^0_{I,1,\cdot} = \mathbf{U}^0_{I,1,\cdot}, \quad \text{Sp } \mathbf{U}^0_{II,1} \subset \mathcal{D}_1, \quad \text{and} \quad \pi_1 \mathbf{U}^0_{II,1} = \mathbf{U}^0_{II,1}.
$$

Then there exists $\tau_2^* > 0$, which depends only on R, such that there exists:

- *a unique $\mathcal{U}_{0,I,\cdot} = \pi_1 \mathcal{U}_{0,I,\cdot} \in D^\sigma_{\tau_2^*}$ solution of*

$$
(3.9)\quad \begin{cases} (\partial_T + v_1 \partial_Z)\mathcal{U}_{0,I,\cdot} = 0, \\ \partial_\tau \mathcal{U}_{0,I,\cdot} + i\dfrac{v_1}{2k_l}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,I,1} + \pi_1 \Lambda(\mathcal{U}_{0,I,\cdot}) = 0, \\ \mathcal{U}_{0,I,\cdot}|_{\tau=T=0} = \mathbf{U}^0_{I,\cdot}, \end{cases}
$$

where $\Lambda(\mathcal{U}_{0,I,\cdot})$ is given by (3.4)*;*

- *a unique $\mathcal{U}^\delta_{0,II,1} = \pi_1 \mathcal{U}^\delta_{0,II,1} \in A^\sigma_{\tau_2^*}$ solution of*

$$
(3.10)\quad \begin{cases} (\partial_T + v_1 \partial_Z)\mathcal{U}^\delta_{0,II,1} = 0, \\ \partial_\tau \mathcal{U}^\delta_{0,II,1} - i\dfrac{v_1}{2D_{z_0}}(\partial_X^2 + \partial_Y^2)\mathcal{U}^\delta_{0,II,1} \\ \qquad + \psi^\delta(D_{z_0})\pi_1 \displaystyle\sum_{k\in Z} F^S(\mathcal{U}_{0,I,2k+1}, \overline{\mathcal{U}_{0,I,2k+1}}, \mathcal{U}^\delta_{0,II,1}) \\ \qquad + \psi^\delta(D_{z_0})\pi_1 f'(\mathcal{U}^\delta_{0,II,1})(\mathcal{U}_{0,I}) + \psi^\delta(D_{z_0})\pi_1 f(\mathcal{U}^\delta_{0,II,1}) = 0 \\ \mathcal{U}^\delta_{0,II,1}|_{\tau=T=0} = \psi^\delta(D_{z_0})\mathbf{U}^0_{II,1}. \end{cases}
$$

Proof. The first equation of (3.9) is automatically solved by looking for $\mathcal{U}_{0,I,\cdot}$ in the form $\mathcal{U}_{0,I,2j+1}(\tau, T, X, Y, Z) = \mathtt{U}_{0,I,2j+1}(\tau, Z - v_1 T, X, Y)$ for all $j \in Z$. System (3.9) then reduces to the Cauchy problem

$$
\begin{cases} \partial_\tau \mathtt{U}_{0,I,\cdot} + i\dfrac{v_1}{2k_l}(\partial_X^2 + \partial_Y^2)\mathtt{U}_{0,I,1} + \pi_1 \Lambda(\mathtt{U}_{0,I,\cdot}) = 0, \\ \mathtt{U}_{0,I,\cdot}|_{\tau=0} = \mathbf{U}^0_{I,\cdot}, \end{cases}
$$

which is easily solved by Picard iterates in $\mathcal{C}([0, \tau_2^*], D^\sigma_0)$ since its linear part defines a unitary group on D^σ_0, which is a Banach algebra.

For the continuous spectrum component, we cannot give an explicit expression of the solution as in Proposition 2.6, because we have to deal with the nonlinearities. However, the presence of nonlinearities is counterbalanced by the simple form of the transport equation. (We have here a common group velocity for all the frequencies.) In order for $\mathcal{U}^\delta_{0,II,1}$ to satisfy this transport equation, we look for it in the form

$$\mathcal{U}^\delta_{0,II,1}(\tau,T,X,Y,Z,t_0,z_0)=\mathtt{U}^\delta_{0,II,1}(\tau,Z-v_1T,X,Y,t_0,z_0),$$

so that (3.10) reduces to the Cauchy problem

$$\begin{cases}\partial_\tau \mathtt{U}^\delta_{0,II,1}-i\dfrac{v_1}{2D_{z_0}}(\partial_X^2+\partial_Y^2)\mathtt{U}^\delta_{0,II,1}\\ \qquad\quad+\psi^\delta(D_{z_0})\pi_1\displaystyle\sum_{k\in Z}F^S(\mathtt{U}_{0,I,2k+1},\overline{\mathtt{U}_{0,I,2k+1}},\mathtt{U}^\delta_{0,II,1})\\ \qquad\quad+\psi^\delta(D_{z_0})\pi_1 f'(\mathtt{U}^\delta_{0,II,1})(\mathtt{U}_{0,I})+\psi^\delta(D_{z_0})\pi_1 f(\mathtt{U}^\delta_{0,II,1})=0\\ \mathtt{U}^\delta_{0,II,1}|_{\tau=0}=\psi^\delta(D_{z_0})\mathbf{U}^0_{II,1}.\end{cases}$$

This Cauchy problem is solved in $B^\sigma_{\tau_2^*}$ by Picard iterates, using estimates similar to those of Lemma 1.5. □

Remark 3.2. (i) With $\mathcal{U}_0$ being given by Proposition 3.3, system (3.7) is then solved by taking $\pi_1\mathcal{U}_{1,II,1}=0$.

(ii) The results of Proposition 3.3 also hold for $\delta=0$, i.e., there exists a unique solution $\mathcal{U}_{0,II}=\mathcal{U}_{0,II,1}=\pi_1\mathcal{U}_{0,II,1}$ to

$$\begin{cases}(\partial_T+v_1\partial_Z)\mathcal{U}_{0,II,1}=0,\\ \partial_\tau\partial_{z_0}\mathcal{U}_{0,II,1}+\dfrac{v_1}{2}(\partial_X^2+\partial_Y^2)\mathcal{U}_{0,II,1}\\ \qquad\quad+\pi_1\displaystyle\sum_{k\in Z}\partial_{z_0}F^S(\mathcal{U}_{0,I,2k+1},\overline{\mathcal{U}_{0,I,2k+1}},\mathcal{U}_{0,II,1})\\ \qquad\quad+\pi_1\partial_{z_0}f'(\mathcal{U}_{0,II,1})(\mathcal{U}_{0,I})+\pi_1\partial_{z_0}f(\mathcal{U}_{0,II,1})=0,\\ \mathcal{U}_{0,II,1}|_{\tau=T=0}=\mathbf{U}^0_{II,1},\end{cases}\tag{3.11}$$

and the convergence property of Proposition 2.13 can easily be extended to the present case.

3.3. Validity of the approximation. We show here that the approximate solution

$$u^\varepsilon(T,X,Y,Z)=\sqrt{\varepsilon}\mathcal{U}^\varepsilon\left(\varepsilon T,T,X,Y,Z,\frac{T}{\varepsilon},\frac{Z}{\varepsilon}\right),$$

with $\mathcal{U}^\varepsilon=\mathcal{U}_0+\varepsilon\mathcal{U}_1+\varepsilon^2\mathcal{U}_2$, is a good approximation of the exact solution of (1.1). As in section 2, the proof cannot be direct because the presence of low frequencies and $\mathcal{L}^{-1}$-regularity are not compatible. But we can mimic the reasoning of section 2 to obtain a stability result. Before stating the theorem, we recall that to any profile $\mathcal{V}\in A^\sigma_{\tau^*}$, we associate the profile $\underline{\mathcal{V}}^\varepsilon$ defined as

$$\underline{\mathcal{V}}^\varepsilon(\tau,X,Y,Z,t_0,z_0):=\mathcal{V}\left(\tau,\frac{\tau}{\varepsilon},X,Y,Z,t_0,z_0\right).$$

THEOREM 3.4. *Suppose the characteristic variety $\mathcal{C}_\mathcal{L}$ is as in Assumption* 2.3.

Let $\mathbf{U}^0 = \mathbf{U}^0_I + \mathbf{U}^0_{II,1} \in A^{s+4}_0$ *such that* $\mathbf{U}^0_I = \sum_{j\in Z} \mathbf{U}^0_{I,2j+1} e^{i(2j+1)\theta}$, *with* $\mathbf{U}^0_{I,\cdot} \in D^{s+4}_0$, *and suppose that*

$$\pi_1 \mathbf{U}^0_{I,\cdot} = \mathbf{U}^0_{I,\cdot}, \quad \text{Sp } \mathbf{U}^0_{II,1} \subset \mathcal{D}_1, \quad \text{and} \quad \pi_1 \mathbf{U}^0_{II,1} = \mathbf{U}^0_{II,1}.$$

Then for $\tau^*_3 = \min\{\tau^*_1, \tau^*_2\}$, *we have the following:*

(i) *The exact solution* $\mathbf{u}^\varepsilon$ *of* (1.1) *exists on* $[0, \tau^*_3/\varepsilon]$ *and can be written* $\mathbf{u}^\varepsilon(T, X, Y, Z) = \sqrt{\varepsilon}\mathbf{U}^\varepsilon(\varepsilon T, X, Y, Z, T/\varepsilon, Z/\varepsilon)$, *with* $\mathbf{U}^\varepsilon = \mathbf{U}^\varepsilon_I + \mathbf{U}^\varepsilon_{II} \in B^{s+4}_{\tau^*_3}$.

(ii) $\mathcal{U}_{0,I,\cdot}$ *is defined in* $D^{s+4}_{\tau^*_3}$ *as the unique solution of* (3.9), *and we define* $\mathcal{U}_{0,I}$ *as* $\mathcal{U}_{0,I} = \sum_{j\in Z} \mathcal{U}_{0,I,2j+1} e^{i(2j+1)\theta}$.

(iii) $\mathcal{U}_{0,II} = \mathcal{U}_{0,II,1}$ *is defined in* $A^{s+4}_{\tau^*_3}$ *as the unique solution of* (3.11).

(iv) *The profile* $\underline{\mathcal{U}}^\varepsilon_0 \in B^s_{\tau^*_3}$ *associated to* $\mathcal{U}_0 = \mathcal{U}_{0,I} + \mathcal{U}_{0,II} \in A^s_{\tau^*_3}$ *approximates the singular equation* (1.5) *in the sense that*

$$\|\mathbf{U}^\varepsilon_I - \underline{\mathcal{U}}^\varepsilon_{0,I}\|_{B^s_{\tau^*_3}} = O(\varepsilon) \quad \text{and} \quad \|\mathbf{U}^\varepsilon_{II} - \underline{\mathcal{U}}^\varepsilon_{0,II}\|_{B^s_{\tau^*_3}} = o(1).$$

(v) *We also have stability of the approximate solution* u^ε_0 *defined with* $\mathcal{U}_0$,

$$\|\mathbf{u}^\varepsilon_I - u^\varepsilon_{0,I}\| = O(\varepsilon^{3/2}) \quad \text{and} \quad \|\mathbf{u}^\varepsilon_{II} - u^\varepsilon_{0,II}\| = o(\sqrt{\varepsilon}),$$

where the norm can be taken either in $\mathcal{C}([0, \frac{\tau^*_3}{\varepsilon}] \times \mathbb{R}^3)^n$ *or in* $\mathcal{C}([0, \frac{\tau^*_3}{\varepsilon}], L^2(\mathbb{R}^3)^n)$.

4. Examples.

4.1. Lasers with large spectrums. As said in the introduction, we want to study the effects due to the fact that certain lasers have frequencies and wavenumbers which dribble around the theoretical value in a range greater than $O(\varepsilon)$ (typically $O(1)$). In order to make a model out of this phenomenon, we add to the theoretical (sinusoidal) oscillations a corrector with a purely continuous spectrum.

To describe the evolution of the electromagnetic field we use Maxwell equations coupled to a response of the medium by the polarization $\mathbf{p}$, which is described by the anharmonic oscillator model [14]. Once nondimensionalized [7], and omitting the divergence-free equations, the system reads

$$(M) \qquad \begin{cases} \partial_T \mathbf{e}^\varepsilon - \mathbf{curl}\, \mathbf{b}^\varepsilon + \dfrac{\sqrt{\gamma_a}}{\varepsilon}\mathbf{g}^\varepsilon & = 0, \\ \partial_T \mathbf{b}^\varepsilon + \mathbf{curl}\, \mathbf{e}^\varepsilon & = 0, \\ \partial_T \mathbf{p}^\varepsilon - \dfrac{\eta_a}{\varepsilon}\mathbf{q}^\varepsilon & = 0, \\ \partial_T \mathbf{q}^\varepsilon - \dfrac{1}{\varepsilon}(\sqrt{\gamma_a}\mathbf{e}^\varepsilon - \eta_a \mathbf{p}^\varepsilon) - \alpha\gamma_a^{3/2}|\mathbf{p}^\varepsilon|^2\mathbf{p}^\varepsilon & = 0. \end{cases}$$

This system (M) is of type (1.1),

$$L^\varepsilon(\partial)\mathbf{u}^\varepsilon + f(\mathbf{u}^\varepsilon) = 0,$$

with

$$\mathbf{u}^\varepsilon = (\mathbf{e}^\varepsilon, \mathbf{b}^\varepsilon, \mathbf{p}^\varepsilon, \mathbf{q}^\varepsilon)^T \in C^{12},$$

and the nonlinearity is of order 3 and reads

$$f(\mathbf{e}^\varepsilon, \mathbf{b}^\varepsilon, \mathbf{p}^\varepsilon, \mathbf{q}^\varepsilon) = (0, 0, 0, \alpha\gamma_a^{3/2}|\mathbf{p}^\varepsilon|^2\mathbf{p}^\varepsilon)^T.$$

Remark 4.1. The fact that the mapping F associated to f is not trilinear as in Assumption 1.2—since it is semilinear in one of its variables—is not important. Indeed, considering $\tilde{\mathbf{u}}^\varepsilon = (\mathbf{u}^\varepsilon, \overline{\mathbf{u}^\varepsilon}) \in C^{24}$ brings us back to this case.

The operator $L^\varepsilon(\partial)$ reads $L^\varepsilon(\partial) = \partial_T + A_1\partial_X + A_2\partial_Y + A_3\partial_Z + L_0/\varepsilon$, with

$$A_1\partial_X + A_2\partial_Y + A_3\partial_Z = \begin{pmatrix} 0 & -\mathbf{curl} & 0 & 0 \\ \mathbf{curl} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$L_0 = \begin{pmatrix} 0 & 0 & 0 & \sqrt{\gamma_a} Id \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\eta_a Id \\ -\sqrt{\gamma_a} Id & 0 & \eta_a Id & 0 \end{pmatrix}.$$

The characteristic variety $\mathcal{C}_\mathcal{L}$ associated to the symbol $\mathcal{L}(\omega, k) = \omega Id + kA_3 + L_0/i$ is defined by the algebraic equation

$$\omega^2(\omega^2 - 1 - \gamma_a)[(\omega^2 - \eta_a^2)(\omega^2 - k^2) - \gamma_a\omega^2]^2 = 0.$$

This characteristic variety has three singular points which are all of abscissa $k = 0$ and ordinate $0; \pm\sqrt{1+\gamma_a}$. If $(\omega_l, -k_l)$ is on a curved sheet of $\mathcal{C}_\mathcal{L}$, i.e., if $(\omega_l^2 - \eta_a^2)(\omega_l^2 - k_l^2) - \gamma_a\omega_l^2 = 0$, then the group velocity $\omega'(k_l)$ is given by

$$\omega'(k_l) = \frac{k_l}{\omega_l} \frac{\omega_l^2 - \eta_a^2}{(\omega_l^2 - k_l^2) + (\omega_l^2 - \eta_a^2) - \gamma_a}, \tag{4.1}$$

while the dispersive factor $\omega''(k_l)$ reads

$$\omega''(k_l) = \omega'(k_l)\frac{\omega_l - \omega'(k_l)}{\omega_l k_l} - 4\frac{\omega_l\omega'(k_l)^2}{k_l}\frac{\omega_l\omega'(k_l) - k_l}{\omega_l^2 - \eta_a^2}. \tag{4.2}$$

Notice that $\mathcal{C}_\mathcal{L}$ contains three plane sheets, so that Assumption 2.1 is not satisfied since these sheets are parallel. However, as mentioned earlier, the divergence-free conditions satisfied by the electromagnetic field allow us to consider that Assumption 2.1 is fulfilled.

We consider initial conditions of the form

$$\mathbf{u}_\varepsilon^0 = \varepsilon^{1/2}\mathbf{U}^0(X, Y, Z, 0, Z/\varepsilon) = \varepsilon^{1/2}(\mathbf{E}^0, \mathbf{B}^0, \mathbf{P}^0, \mathbf{Q}^0)(X, Y, Z, 0, Z/\varepsilon),$$

where $\mathbf{U}^0$ is written $\mathbf{U}^0(X, Y, Z, t_0, z_0) = \mathbf{U}_{I,1}^0(X, Y, Z)e^{i\theta} + \text{c.c.} + \mathbf{U}_{II}^0(X, Y, Z, t_0, z_0)$, with the component $\mathbf{U}_{II}^0$ having a purely continuous spectrum. As mentioned above, $\mathbf{U}_{I,1}^0$ corresponds to the usual (small-spectrum) laser, i.e., the laser with time-space wavenumber equal to $(\omega_l, -k_l)$, while $\mathbf{U}_{II}^0$ corresponds to the large dribbling.

Moreover, the initial conditions are polarized,

$$\pi(\omega_l, -k_l)\mathbf{U}_{I,1}^0 = \mathbf{U}_{I,1}^0 \quad \text{and} \quad \pi(D_{t_0,z_0})\mathbf{U}_{II}^0 = \mathbf{U}_{II}^0.$$

The solution of diffractive optics reads

$$\begin{aligned} u^\varepsilon(T, X, Y, Z) = \varepsilon^{1/2}(&\mathcal{U}_{0,I,1}(\varepsilon T, T, X, Y, Z)e^{i(\omega_l\frac{T}{\varepsilon} - k_l\frac{Z}{\varepsilon})} + \text{c.c.} \\ &+ \mathcal{U}_{0,II}\left(\varepsilon T, T, X, Y, Z, \frac{T}{\varepsilon}, \frac{Z}{\varepsilon}\right), \end{aligned}$$

and the results of section 2 state that the profile $\mathcal{U}_{0,I,1}$ is given by

$$\partial_\tau \mathcal{U}_{0,I,1} + i\frac{\omega'(k_l)}{2k_l}(\partial_X^2+\partial_Y^2)\mathcal{U}_{0,I,1} + i\frac{\omega''(k_l)}{2}\partial_Z^2\mathcal{U}_{0,I,1} + \pi(\omega_l,-k_l)f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}}) = 0,$$

with here

$$f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}}) = \left(0,0,0,\alpha\gamma_a^{3/2}\left(2|\overrightarrow{\mathcal{P}}_{0,I,1}|^2\overrightarrow{\mathcal{P}}_{0,I,1} + (\overrightarrow{\mathcal{P}}_{0,I,1}.\overrightarrow{\mathcal{P}}_{0,I,1})\overline{\overrightarrow{\mathcal{P}}_{0,I,1}}\right)\right)^T$$

and ω' and ω'' given by (4.1)–(4.2).

As the waves which we consider here propagate along (OZ) with a wavenumber $\overrightarrow{k_l} = (0,0,k_l)$, the electric field is polarized on the plane (OXY). We can assume that it is polarized along (OX), i.e., $\overrightarrow{\mathcal{E}}_{0,I,1} = (\mathcal{E}_{0,I,1},0,0)^T$. Since $\pi(\omega_l,-k_l)\mathcal{U}_{0,I,1} = \mathcal{U}_{0,I,1}$, one therefore has $\overrightarrow{\mathcal{P}}_{0,I,1} = \eta_a\chi(\omega_l)\overrightarrow{\mathcal{E}}_{0,I,1}$, where the dielectric susceptibility $\chi(\omega_l)$ is given by

$$\chi(\omega_l) = \frac{\sqrt{\gamma_a}}{\eta_a^2-\omega_l^2}.$$

The nonlinearity therefore reads

$$f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}}) = (0,0,0,3\alpha\gamma_a^{3/2}\eta_a^3\chi(\omega_l)^3|\mathcal{E}_{0,I,1}|^2\mathcal{E}_{0,I,1},0,0)^T.$$

In order to obtain the evolution equation on $\mathcal{E}_{0,I,1}$, one needs to compute the nonlinearity $\pi(\omega_l,-k_l)f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}})$ (in fact, computing its first component is enough). For all vectors $a = (a_1,0,0)^T \in C^3$, one has

$$\pi(\omega_l,-k_l)\begin{pmatrix}0\\0\\0\\a\end{pmatrix} = -\frac{i\omega_l\sqrt{\gamma_a}}{\eta_a^2-\omega_l^2}\begin{pmatrix}\frac{a_1}{N^2}\\.\\.\\.\end{pmatrix}$$

with

$$\begin{aligned} N^2 &= 1+\frac{k_l^2}{\omega_l^2}+\eta_a^2\chi^2(\omega_l)+\omega_l^2\chi^2(\omega_l)\\ &= \sqrt{\gamma_a}\left(\frac{k_l^2+\omega_l^2}{k_l^2-\omega_l^2}+\frac{\eta_a^2+\omega_l^2}{\eta_a^2-\omega_l^2}\right)\chi(\omega_l)\\ &:= \sqrt{\gamma_a}\beta(\omega_l,k_l)\chi(\omega_l). \end{aligned}$$

We thus find

$$\pi(\omega_l,-k_l)f'(\mathcal{U}_{0,I,1})(\overline{\mathcal{U}_{0,I,1}}) = -3i\alpha\gamma_a\eta_a^3\frac{\omega_l}{\beta(\omega_l,k_l)}\chi(\omega_l)^3(|\mathcal{E}_{0,I,1}|^2\mathcal{E}_{0,I,1},\ldots),$$

and the evolution equation on $\mathcal{E}_{0,I,1}$ is therefore

$$\begin{aligned} &\partial_\tau\mathcal{E}_{0,I,1} + i\frac{\omega'(k_l)}{2k_l}(\partial_X^2+\partial_Y^2)\mathcal{E}_{0,I,1} + i\frac{\omega''(k_l)}{2}\partial_Z^2\mathcal{E}_{0,I,1}\\ &= -3i\alpha\gamma_a\eta_a^3\frac{\omega_l}{\beta(\omega_l,k_l)}\chi(\omega_l)^3|\mathcal{E}_{0,I,1}|^2\mathcal{E}_{0,I,1}. \end{aligned}$$

The purely continuous spectrum component of $\mathcal{U}_0$ is found solving

$$\partial_\tau \partial_{z_0} \mathcal{U}_{0,II} - \frac{\omega'(D_{z_0})}{2}(\partial_X^2 + \partial_Y^2)\mathcal{U}_{0,II} - \frac{D_{z_0}\omega''(D_{z_0})}{2}\partial_Z^2 \mathcal{U}_{0,II} = 0.$$

Supposing as above that $\mathcal{E}_{0,II}$ is polarized along (OX), i.e., $\overrightarrow{\mathcal{E}}_{0,II} = (\mathcal{E}_{0,II}, 0, 0)^T$, we obtain

$$\partial_\tau \partial_{z_0} \mathcal{E}_{0,II} - \frac{\omega'(D_{z_0})}{2}(\partial_X^2 + \partial_Y^2)\mathcal{E}_{0,II} - \frac{D_{z_0}\omega''(D_{z_0})}{2}\partial_Z^2 \mathcal{E}_{0,II} = 0.$$

Remark 4.2. A usual direct computation of this equation without the help of the general results proved above would have led to a nonlinear equation. It was not obvious a priori that all the nonlinear terms would be negligible.

4.2. Short pulses. The second application we study concerns short-pulse lasers. Normally, the length of the pulse of a laser is long enough to contain many oscillations so that the phenomena considered can be described well enough by knowing the evolution of the envelope of these oscillations. For ultrashort pulses, this is no longer true (see Figure 1) since there may even be less than an oscillation. The profile we use to make a model of this phenomenon therefore has only a purely continuous spectrum component: the sinusoidal one (discrete spectrum) does not have time to appear. More precisely, we consider an initial condition for (M) of the form

$$\mathbf{u}_\varepsilon^0 = \varepsilon^{1/2}\mathbf{U}^0(X, Y, Z, 0, Z/\varepsilon) = \varepsilon^{1/2}(\mathbf{E}^0, \mathbf{B}^0, \mathbf{P}^0, \mathbf{Q}^0)(X, Y, Z, 0, Z/\varepsilon),$$

where $\mathbf{U}^0 \in A_0^s$ has a purely continuous spectrum. Moreover the initial condition is polarized as follows:

$$\pi(D_{t_0,z_0})\mathbf{U}^0 = \mathbf{U}^0.$$

In accordance with the results of section 2, the profile $\mathcal{U}_0 = \mathcal{U}_{0,II}$ of the approximate solution is found solving

$$\partial_\tau \partial_{z_0} \mathcal{U}_0 - \frac{\omega'(D_{z_0})}{2}(\partial_X^2 + \partial_Y^2)\mathcal{U}_0 - \frac{D_{z_0}\omega''(D_{z_0})}{2}\partial_Z^2 \mathcal{U}_0 = 0.$$

Considering the same model (M) as in the previous section and using the same notation thus yield the following equation for the nonzero component $\mathcal{E}_0$ of the electric field:

$$\partial_\tau \partial_{z_0} \mathcal{E}_0 - \frac{\omega'(D_{z_0})}{2}(\partial_X^2 + \partial_Y^2)\mathcal{E}_0 - \frac{D_{z_0}\omega''(D_{z_0})}{2}\partial_Z^2 \mathcal{E}_0 = 0,$$

where ω' and ω'' are given by (4.1)–(4.2).

Remark 4.3. Here again the fact that we would obtain a linear equation is not obvious. We also point out the fact that our framework allows us to find this equation in the physical *dispersive* case.

In a nondispersive framework, the nonlinearities would not have vanished. Assuming that the spectrum of $\mathcal{U}_0 = \mathcal{U}_{0,II}$ is located on the line $\mathcal{D}_1$ of $\mathcal{C}_\mathcal{L}$ to which $(\omega_l, -k_l)$ belongs, we would find

$$\partial_\tau \partial_{z_0} \mathcal{U}_0 + \frac{v_1}{2}(\partial_X^2 + \partial_Y^2)\mathcal{U}_0 + \pi_1 \partial_{z_0} f(\mathcal{U}_0) = 0,$$

where $-v_1$ is the slope of D_1. This is Alterman and Rauch's equation [1], [2], [3].

5. Weakly dispersive case. We have seen in the previous sections that there is a radical difference between the behavior of the approximate solution of the dispersive case and that of the nondispersive case. In the former, the evolution of the continuous spectrum mode is uncoupled with the discrete spectrum mode and is linear; in the latter, it is coupled and nonlinear.

This behavioral gap means that the estimate $o(1)$ of Theorem 2.14 is somewhat critical for weakly dispersive systems.We propose in this section a few hints to improve this result.

In terms of BKW strategy, solving (2.23) or the following equations is equivalent:

$$\left\{\begin{array}{l} \pi(D_{t_0},D_{z_0})\mathcal{U}_{0,II}=\mathcal{U}_{0,II},\\ (\partial_T+\omega'(k_l)\partial_Z)\mathcal{U}_{0,II}=0,\\ \partial_\tau\mathcal{U}_{0,II}+i\dfrac{\omega'(D_{z_0})}{2D_{z_0}}(\partial_X^2+\partial_Y^2)\mathcal{U}_{0,II}+\dfrac{-\omega'(k_l)-\omega'(D_{z_0})}{\varepsilon}\partial_Z\mathcal{U}_{0,II}\\ \qquad +i\dfrac{\omega''(D_{z_0})}{2}\partial_Z^2\mathcal{U}_{0,II}=6\pi(D_{t_0},D_{z_0})F^S(\mathcal{U}_{0,I,1},\overline{\mathcal{U}_{0,I,1}},\mathcal{U}_{0,II}). \end{array}\right. \tag{5.1}$$

Note that the nonlinearity has been simplified according to the results of Lemmas 2.5 and 3.2, and that here $\pi_1(D_{t_0},D_{z_0})\mathcal{U}_{1,II}=0$.

In general, the Fourier multiplier $\frac{-\omega'(k_l)-\omega'(D_{z_0})}{\varepsilon}$ is of size $O(1/\varepsilon)$ and solving (2.23) is much more relevant than solving (5.1) since this former system does not involve ε. However, if the model considered is weakly dispersive in the sense that one can choose ω_0' in such a way that $\frac{-\omega'(k_l)-\omega'(D_{z_0})}{\varepsilon}=O(1)$, system (5.1) becomes more interesting because the transport equation is the same for all frequencies, as in the nondispersive case. Since it is also reasonable to suppose that such weakly dispersive systems also satisfy $\omega''(D_{z_0})=O(\varepsilon)$, system (5.1) is equivalent in terms of BKW strategy to

$$\left\{\begin{array}{l} \pi(D_{t_0},D_{z_0})\mathcal{U}_{0,II}=\mathcal{U}_{0,II},\\ (\partial_T+\omega'(k_l)\partial_Z)\mathcal{U}_{0,II}=0,\\ \partial_\tau\partial_{z_0}\mathcal{U}_{0,II}+\dfrac{\omega'(k_l)}{2}(\partial_X^2+\partial_Y^2)\mathcal{U}_{0,II}+\dfrac{-\omega'(k_l)-\omega'(D_{z_0})}{\varepsilon}\partial_Z\partial_{z_0}\mathcal{U}_{0,II}\\ \qquad =6\pi(D_{t_0},D_{z_0})\partial_{z_0}F^S(\mathcal{U}_{0,I,1},\overline{\mathcal{U}_{0,I,1}},\mathcal{U}_{0,II}), \end{array}\right. \tag{5.2}$$

where we also have differentiated the last equation with respect to z_0.

Therefore, in the weakly dispersive case, one can obtain profile equations for $\mathcal{U}_{0,II}$ which are linear but coupled with the evolution equation of the discrete spectrum mode $\mathcal{U}_{0,I,1}$. System (5.2) thus provides an intermediate model between the dispersive case of section 2 and the nondispersive case of section 3, and hence partially fills the behavioral gap between these two situations.

Example. With the same notation as in section 4.1, the equation for the electric field $\mathcal{E}_{0,II}$ associated to a large-spectrum laser reads

$$\begin{aligned}\partial_\tau\partial_{z_0}\mathcal{E}_{0,II}+\frac{\omega'(k_l)}{2}(\partial_X^2+\partial_Y^2)\mathcal{E}_{0,II}+\frac{-\omega'(k_l)-\omega'(D_{z_0})}{\varepsilon}\partial_Z\partial_{z_0}\mathcal{E}_{0,II}\\ =-6i\alpha\gamma_a\eta_a^3\chi_{\omega_l}^2\frac{D_{t_0}\chi(D_{z_0})}{\beta(D_{t_0},D_{z_0})}|\mathcal{E}_{0,I,1}|^2\mathcal{E}_{0,II},\end{aligned}$$

which is still linear but coupled with $\mathcal{E}_{0,I,1}$. We recall that by weakly dispersive we mean that $\frac{-\omega'(k_l)-\omega'(D_{z_0})}{\varepsilon}=O(1)$, and that we must simultaneously solve $(\partial_T+\omega'(k_l)\partial_Z)\mathcal{E}_{0,II}=0$.

Acknowledgments. The authors want to give warm thanks to T. Colin and G. Gallice for fruitful discussions. They are also grateful to J. Rauch and E. Dumas for their remarks and advice on a previous version of this work.

REFERENCES

[1] D. Alterman, *Diffractive Nonlinear Geometric Optics for Short Pulses*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 1999.
[2] D. Alterman and J. Rauch, *Diffractive short pulse asymptotics for nonlinear wave equations*, Phys. Lett. A, 264 (2000), pp. 390–395.
[3] D. Alterman and J. Rauch, *Diffractive nonlinear geometric optics for short pulses*, to appear.
[4] D. Alterman and J. Rauch, *Nonlinear geometric optics for short pulses*, J. Differential Equations, 178 (2002), pp. 437–465.
[5] R. Benedetti and J.-J. Risler, *Real algebraic and semi-algebraic sets*, Actualités Math., Hermann, Paris, 1990.
[6] J. Bochnak, M. Coste, and M.-F. Roy, *Géométrie algébrique réelle*, Ergeb. Math. Grenzgeb. (3) 12, Springer-Verlag, Berlin, 1987.
[7] P. Donnat, *Quelques contributions en optique non linéaire*, Ph.D. thesis, École Polytechnique, Palaiseau, France, 1994.
[8] P. Donnat, J.-L. Joly, G. Metivier, and J. Rauch, *Diffractive nonlinear optics*, in Sémin. Équ. Dériv. Partielles Exp. XVII, École Polytechnique, Palaiseau, France, 1996.
[9] J.-L. Joly, *Sur la propagation des oscillations par un système hyperbolique semi-linéaire en dimension* 1 *d'espace*, C. R. Acad. Sci. Paris Sér. I Math., 296 (1983), pp. 669–672.
[10] J.-L. Joly, G. Métivier, and J. Rauch, *Diffractive nonlinear geometric optics with rectification*, Indiana Univ. Math. J., 47 (1998), pp. 1167–1241.
[11] D. Lannes, *Dispersive effects for nonlinear geometrical optics with rectification*, Asymptot. Anal., 18 (1998), pp. 111–146.
[12] D. Lannes, *Nonlinear geometrical optics for oscillatory wave trains with continuous oscillatory spectrum*, Adv. Differential Equations, 6 (2001), pp. 731–768.
[13] O. Morice, *Obtention de l'équation de schrödinger modifiée*, personal communication, 1999.
[14] A. Newell and J. Moloney, *Nonlinear Optics*, Addison–Wesley, Reading, MA, 1992.
[15] J. Rauch, *Nonlinear pulse propagation*, in Journées "Équations aux Dérivées Partielles" (Plestin-les-grèves, 2001), Exp. XI, École Polytechnique, Palaiseau, France, 2001.
[16] A. Yoshikawa, *Solutions containing a large parameter of a quasi-linear hyperbolic system of equations and their nonlinear geometric optics approximation*, Trans. Amer. Math. Soc., 340 (1993), pp. 103–126.
[17] A. Yoshikawa, *Asymptotic expansions of the solutions to a class of quasilinear hyperbolic initial value problems*, J. Math. Soc. Japan, 47 (1995), pp. 227–252.

SIAM J. MATH. ANAL.
Vol. 34, No. 3, pp. 675–699

THE TRANSITION FROM ZELDOVICH–VON NEUMANN–DORING TO CHAPMAN–JOUGUET THEORIES FOR A NONCONVEX SCALAR COMBUSTION MODEL*

JIEQUAN LI† AND PENG ZHANG‡

Abstract. We study the transition from the Zeldovich–von Neumann–Doring (ZND) theory to the Chapman–Jouguet (CJ) theory as the reaction rate tends to infinity for a nonconvex scalar combustion model. The Riemann solution of the nonconvex ZND combustion model is constructed, and the limit of solutions as the reaction rate goes to infinity is investigated. We classify the reaction solutions of the ZND combustion model as detonation and deflagration waves according to the essential difference that the former contains the von Neumann spike but the latter does not. Based on the analysis of this limit, we propose a set of entropy conditions for combustion and noncombustion waves to the nonconvex CJ combustion model, which is the indispensable preparation for the study of multidimensional combustion problems.

Key words. Zeldovich–von Neumann–Doring theory, Chapman–Jouguet theory, detonation, deflagration, von Neumann spike, reaction entropy condition

AMS subject classifications. Primary, 35L60, 35L67; Secondary, 76L05, 80A25

PII. S0036141001386751

1. Introduction.

The Zeldovich–von Neumann–Doring (ZND) theory and the Chapman–Jouguet (CJ) theory play an important role in gas dynamics combustion theory. The former describes the combustible gas with a finite reaction rate and the latter with an infinite reaction rate or, equivalently, the infinitely thin reaction region. Formally the CJ theory is regarded as the limit of the ZND theory as the reaction rate tends to infinity [1]. Interesting discussions on this transition can be found in [3, 16]. The Riemann problem for the CJ gas dynamic combustion is constructively solved in [22] to display rich combustion wave patterns satisfying the so-called geometrical entropy conditions. Yet, the study of combustion waves of gas dynamics based on the ZND theory is notoriously complex and difficult, and far from being complete [21]. This motivates us to consider simpler combustion models.

In [2, 14], Fickett and Majda independently proposed a simplified model to study combustion waves, as the Burgers equation models in gas dynamics,

$$\begin{aligned} &(u+qz)_t + f(u)_x = \mu u_{xx}, \\ &z_t + k\phi(u)z = 0, \end{aligned} \tag{1.1}$$

where $x \in \mathbb{R}$, $t > 0$. This model corresponds to the ZND model in gas dynamics. The first equation resembles the conservation law of energy in gas dynamics, leading to nonlinear phenomena such as shocks; the second is the reaction equation, which

*Received by the editors March 25, 2001; accepted for publication (in revised form) May 22, 2002; published electronically January 7, 2003. This research was partially supported by the NNSF of China.

http://www.siam.org/journals/sima/34-3/38675.html

†Department of Mathematics, Capital Normal University, Beijing, 100037, People's Republic of China (jiequan@mail.cnu.edu.cn). This author was supported by a Golda Meir Fellowship awarded by the Hebrew University of Jerusalem.

‡Beijing Information Technology Institute, Beijing, 100101, People's Republic of China (pzhang@ht.rol.cn.net). This work was completed during this author's stay in the Institute of Mathematics, Hebrew University of Jerusalem.

may give rise to a linear discontinuity in the solution. The dependent variable u is a lumped quantity, representing density, velocity, or temperature, but not the chemical binding energy, which is represented by q. z is the percentage of unburned gas, k is the reaction rate, and $\phi(u)$ is the standard Heaviside function. We take the ignition point $u = 0$ for simplicity. That is, the material begins to burn just as the temperature becomes higher than zero. Using this model, combustion problems were extensively investigated, such as the stability and asymptotic behavior of combustion waves (see [5, 9, 10, 11] and the references therein). The well-posedness of the general Cauchy problem and the zero viscosity limit of (1.1) was studied in [5]. Ying and Teng [17] studied the Riemann solution of (1.1) at $\mu = 0$ and obtained the limit of the solution as k tends to infinity and defined the limit function as the solution of the Riemann problem for the corresponding CJ model

$$\begin{aligned} &(u+qz)_t + f(u)_x = 0, \\ &z(x,t) = \begin{cases} z(x,0) & \text{if } \max\limits_{0\le\tau\le t} u(x,\tau) \le 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{1.2}$$

In [13], Liu and Zhang extracted from the properties of CJ solutions in [17] to propose a set of entropy conditions, under which the existence and uniqueness of the Riemann solutions were shown constructively and the ignition and extinction problems were investigated as well. A slightly different model was considered in [7, 8, 12, 15]. All of these results were obtained under the assumption that $f(u)$ is strictly convex.

Based on these results in one dimension, one would naturally wish to study combustion phenomena in several dimensions. The generalization of (1.1) may be a reasonable trial in this aspect, for which the theory of well-posedness of a multidimensional scalar ZND model was established in [6]. As is well known, the fluxes must be nonconvex in some directions at this moment [4]. Therefore, it is interesting to investigate scalar combustion models with nonconvex fluxes $f(u)$ so as to make a good preparation for the study of structure of multidimensional combustion waves.

To this end, the Riemann problem for (1.2) with nonconvex fluxes $f(u)$ was solved constructively in [18] under the entropy restriction that mimics those in [13] and generalizes the classical Oleinik entropy condition for scalar conservation laws. A crucial issue here is just how to propose and justify entropy conditions to single out physically admissible solutions. There are some plausible ways: One is, most naturally, to consider the viscosity vanishing limit for corresponding models like (1.1) with nonconvex fluxes, which is found to be a very difficult issue, even only to discuss traveling wave solutions. Another is, as in the classical gas dynamics combustion theory [1], to investigate the limit of the nonviscous Fickett–Majda model (1.1) as the reaction rate k goes to infinity. This is what we do in this paper.

The self-similar combustion model corresponding to (1.1) with sufficiently large reaction rates reads

$$\begin{aligned} &(u+qz)_t + f(u)_x = 0, \\ &z_t + \tfrac{k}{t}\phi(u)z = 0, \end{aligned} \tag{1.3}$$

where $f(u)$ is nonconvex. The second author began to study this model and announced the very partial results in [19]; the Riemann problem was discussed there for some cases and the large reaction rate limit was taken into account to get the associated Riemann solutions of the corresponding CJ model (1.2). However, this result is far

from understanding the entropy combustion solutions which we clarify below. As pointed out in [13], even the Riemann solution to (1.3) is not unique, provided that it is restricted by the classical Oleinik-type entropy condition (2.7). Motivated by the physical consideration that the temperature in the combustion wave front is as high as possible or, equivalently, the propagation speed of combustion wave front is as small as possible, we propose in this paper a reaction entropy condition to guarantee the uniqueness of solutions, under which the Riemann solution to (1.3) is uniquely constructed. Since we focus our attention on the transition from the ZND theory to the CJ theory as the reaction rate k goes to infinity, we mainly display the structure of solutions with the large reaction rate k. Our main contribution is to clarify the reaction solution as detonation and deflagration waves in accord with the essential difference that the former has a von Neumann spike in the reaction region while the latter does not, although it was already noticed earlier, e.g., in [1, 22, 16, 19]. Indeed, the temperature (the lumped variable u) is not monotone for the former but is monotone increasing for the latter. Then we study the limit of solutions by letting the reaction rate k tends to infinity. It is found that although both of these combustion waves in the limit become jump-ups, they still inherit the above intrinsic difference. Finally we formulate a set of entropy conditions based on the analysis of limit solutions, which enables us to greatly improve the result in [18] to get the unique entropy solution of (1.2) with nonconvex flux $f(u)$. Indeed, this entropy condition can be used to justifiably construct two-dimensional Riemann solutions for the CJ combustion model associated with (1.2); see [20].

The rest of this paper consists of three parts. In section 2, we give some preliminaries containing the general property of smooth solutions, the Rankine–Hugoniot jump conditions of combustion waves, and the reaction entropy condition. The Riemann solutions of (1.3) are constructed in section 3 for two typically distinct fluxes with just one inflection point. The limit behavior of solutions is also studied as the reaction rate goes to infinity. We propose the entropy condition from the limit behavior of solutions in the preceding sections for the CJ nonconvex combustion model (1.2) in section 4.

2. Self-similar solutions to the ZND model. This section serves as a preliminary for the forthcoming sections. We will discuss the general properties of smooth solutions, the Rankine–Hugoniot jump conditions of combustion waves, and the reaction entropy condition.

We begin by considering the Riemann problem for (1.3) with initial data

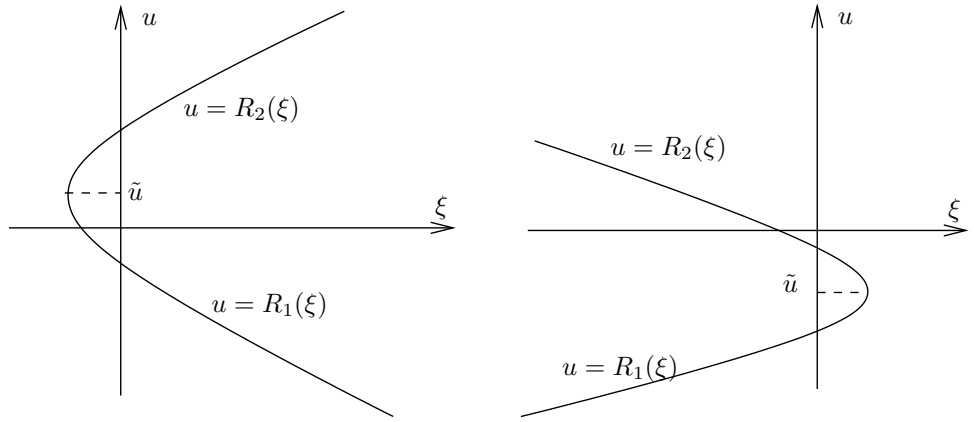
$$(u,z)|_{t=0} = \begin{cases} (u^-, z^-), & x < 0, \\ (u^+, z^+), & x > 0, \end{cases} \tag{2.1}$$

where (u^-, z^-) and (u^+, z^+) are two different states. For the self-similar solutions $(u(\xi), z(\xi))$, $\xi = x/t$, the Riemann problem for (1.3) becomes a boundary value problem with boundary values at infinity,

$$\begin{aligned} &(f'(u) - \xi)\frac{du}{d\xi} = qk\phi(u)z, \\ &\xi\frac{dz}{d\xi} = k\phi(u)z, \end{aligned} \tag{2.2}$$

and

$$(u,z)|_{\xi=\pm\infty} = (u^\pm, z^\pm). \tag{2.3}$$



(a) The branches with assumption (A_1). (b) The branches with assumption (A_2).

FIG. 2.1. *The branches of the inverse function of* $\xi = f'(u)$.

Without loss of generality, we consider this Riemann problem with the data

$$z^- = 1,\ z^+ = 0; \qquad u^- \le 0 < u^+. \tag{2.4}$$

The left state $(u^-, 1)$ is unburned and the right state $(u^+, 0)$ is burnt. For clarity of presentation, we restrict $f(u)$ to be the following two typical cases since the general case for $f(u)$ with a finite number of isolated inflection points is treated similarly without substantial difficulties.

ASSUMPTIONS.

(A_1) *$f(u)$ has one inflection point $\tilde{u}$ and $f'(\pm\infty) = +\infty$;*

(A_2) *$f(u)$ has one inflection point $\tilde{u}$ and $f'(\pm\infty) = -\infty$.*

These assumptions are reasonable in the application to the multidimensional generalization. Consider a two-dimensional counterpart of (1.1) (see [6]),

$$\begin{aligned} &(u + qz)_t + f(u)_x + g(u)_y = \epsilon \Delta u, \\ &z_t + k\phi(u)z = 0, \end{aligned} \tag{2.5}$$

where Δ is the Laplacian, the fluxes $f(u)$ and $g(u)$ satisfy that $f''(u) > 0$, $g''(u) > 0$, and $(f''(u)/g''(u))' > 0$. It is easily shown that for any given direction $(\mu, \nu) \in S^1$, the directional flux $F(u; \mu, \nu) = \mu f(u) + \nu g(u)$ has at most one inflection point.

For each case of these assumptions, the inverse function of $\xi = f'(u)$ has two branches, denoted by $R_1(\xi)$ and $R_2(\xi)$, where $R_1(\xi) < R_2(\xi)$. More specifically, with assumption (A_1), the flux $f(u)$ is convex when $u > \tilde{u}$, while it is concave when $u < \tilde{u}$. Therefore, $u = R_1(\xi)$ is defined via $\xi = f'(u)$ as $u < \tilde{u}$ and $u = R_2(\xi)$ via $\xi = f'(u)$ as $u > \tilde{u}$. Thus, $u = R_1(\xi)$ is decreasing while $u = R_2(\xi)$ is increasing. For (A_2), we have converse statements. See Figure 2.1 for the graphs of $R_1(\xi)$ and $R_2(\xi)$.

Thus, the smooth solutions of (2.2) are in the following:

(1) Constant states, $(u, z) =$ (constant, constant).

(2) $(u, z) = (R_i(\xi),$ constant), where $i = 1$ or 2.

(3) $z(\xi) = |\frac{\xi}{\eta}|^k$, where η is an arbitrary constant, $u(\xi) > 0$ satisfies

$$(f'(u) - \xi)\frac{du}{d\xi} = qk\left|\frac{\xi}{\eta}\right|^k.$$

We now turn to discuss discontinuous solutions. At this moment, we have to understand (2.2) in the sense of distributions. Let $(u(\xi), z(\xi))$ be a piecewise smooth solution with a discontinuity point at $\xi = \sigma$. Then we get the Rankine–Hugoniot jump condition

$$\begin{aligned} &-\sigma[u] + [f] = 0, \\ &\sigma[z] = 0. \end{aligned} \tag{2.6}$$

Throughout this paper, we fix the notation $[u]$ to be the jump of u across $\xi = \sigma$, etc. Then the Rankine–Hugoniot condition (2.6) provides two possibilities: (i) if $[z] \neq 0$, then $\sigma = 0$ and $[f] = 0$; (ii) if $[z] = 0$, then $\sigma = \frac{[f]}{[u]}$. The first corresponds to a slip line and the second a shock wave, provided that it satisfies the following Oleinik-type entropy condition:

$$\frac{f(u) - f(u_l)}{u - u_l} \geq \frac{f(u_r) - f(u_l)}{u_r - u_l} \quad \text{for } (u - u_l)(u - u_r) \leq 0. \tag{2.7}$$

The pair (u, z) is an entropy solution of (2.2) and (2.4) if the equations are satisfied at smooth points in the classical sense and the requirement of Oleinik-type entropy condition (2.7) is met at discontinuity points. This solution has the following property.

LEMMA 2.1. *Let $(u(\xi), z(\xi))$ be an entropy solution of* (2.2) *and* (2.4). *Then there exists $\eta \in (-\infty, 0]$ such that $z(\xi)$ has the structure*

$$z(\xi) = \begin{cases} 1, & \xi < \eta, \\ (\frac{\xi}{\eta})^k, & \eta \leq \xi \leq 0, \\ 0, & 0 < \xi. \end{cases} \tag{2.8}$$

Proof. With the above arguments and the Oleinik-type entropy condition (2.7), $z(\xi)$ consists of some constants and functions with the form $|\frac{\xi}{\eta}|^k$. The only possible discontinuity point of $z(\xi)$ is $\xi = 0$. Since $z(+\infty) = 0$, $z(\xi) \equiv 0$ for $\xi > 0$. Note that $z(-\infty) = 1$. We assert $z(\xi) \equiv 1$ in a neighborhood of negative infinity. If $z(\xi) \equiv 1$ for all $\xi < 0$, then $z(\xi)$ has the structure of (2.8) with $\eta = 0$. Otherwise, there exists a constant $\eta < 0$ satisfying the Rankine–Hugoniot condition $\sigma = \eta = (f(u(\eta - 0)) - f(u(\eta + 0))/(u(\eta - 0)) - u(\eta + 0))$, where u undergoes a jump (shock wave). Since z is continuous there, we conclude that $z(\eta) = 1$, and so $z(\xi) = (\xi/\eta)^k$ for $\xi \in (\eta, 0)$ by (2.2). Thus $z(\xi)$ has three different stages as expressed in (2.8). ∎

Lemma 2.1 gives the following corollary.

COROLLARY 2.2. *The entropy solution $(u(\xi), z(\xi))$ of* (2.2) *and* (2.4) *has the structure*

$$(u(\xi), z(\xi)) = \begin{cases} (A(\xi), 1), & -\infty < \xi < \eta, \\ (B(\xi), (\frac{\xi}{\eta})^k), & \eta \leq \xi \leq 0, \\ (C(\xi), 0), & 0 < \xi < +\infty, \end{cases} \tag{2.9}$$

where A, B, and C satisfy the following equations at smooth points:

$$\begin{aligned} &(f'(A)-\xi)\frac{dA}{d\xi}=0, \quad \xi\in(-\infty,\eta),\\ &A(-\infty)=u^-, \quad A(\eta-0)\le 0, \end{aligned} \tag{2.10}$$

$$\begin{aligned} &(f'(B)-\xi)\frac{dB}{d\xi}=qk\left(\frac{\xi}{\eta}\right)^k, \quad \xi\in[\eta,0],\\ &B(\xi)\ge 0, \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} &(f'(C)-\xi)\frac{dC}{d\xi}=0, \quad \xi\in(0,+\infty),\\ &C(+\infty)=u^+. \end{aligned} \tag{2.12}$$

The states $A(\xi)$, $B(\xi)$, and $C(\xi)$ are called the unburned, burning, and burnt parts of $u(\xi)$, respectively. The pair (u,z) is a combustion solution if it contains a burning part; otherwise, it is a noncombustion solution.

Analogous to [13], where $f(u)$ is convex (or concave), it can be shown that the entropy solution of (2.2) and (2.4) is not unique for some binding energy q and initial data (2.4) even though we impose the requirement of the entropy condition (2.7). Motivated by the physical consideration that the temperature in the wave front is as high as possible or, equivalently, the propagation speed of combustion wave front is as small as possible (cf. [22]), we propose the following *reaction entropy condition* to guarantee the uniqueness of solutions, in addition to the Oleinik-type entropy condition (2.7).

Reaction entropy condition (BE). *If the Riemann problem of* (2.2) *and* (2.4) *has several entropy solutions, we choose a solution so that its speed η in the wave front of the burning part achieves the absolute minimum value.*

A solution (u,z) is *admissible* if it satisfies both the Oleinik-type entropy condition (2.7) and the reaction entropy condition (BE). We call (2.7) and (BE) together as the entropy condition of the reaction solution of (2.2) and (2.4).

3. The entropy solution and the limit of the infinite reaction rate. In this section, we seek the entropy solution to (2.2) and (2.4) when the flux $f(u)$ satisfies the assumptions in the last section. We not only prove the solvability of this problem but display the explicit structure of solutions as well. Furthermore, we consider the limit behavior of entropy solutions as the reaction rate k goes to infinity. We achieve our goals through two cases according to the shape of flux $f(u)$ in the assumptions in the preceding section.

3.1. The solution of (2.2) and (2.4) when $f(u)$ satisfies (A_1). We consider the solution of (2.2) and (2.4) when $f(u)$ has only one inflection point and the slope at infinity is positive infinity. The main results are stated in Theorems 3.7–3.9. We investigate this problem using three cases depending on the position of u^+.

Let $\tilde{u}$ be the inflection point of $f(u)$. If $f'(\tilde{u})\ge 0$, then $f(u)$ is a monotone increasing function. It is easy to verify that the Riemann problem of (2.2) and (2.4) has a unique noncombustion solution (cf. [13]). Therefore, the admissible solution exists and is unique subject to the above entropy conditions. In the following, we

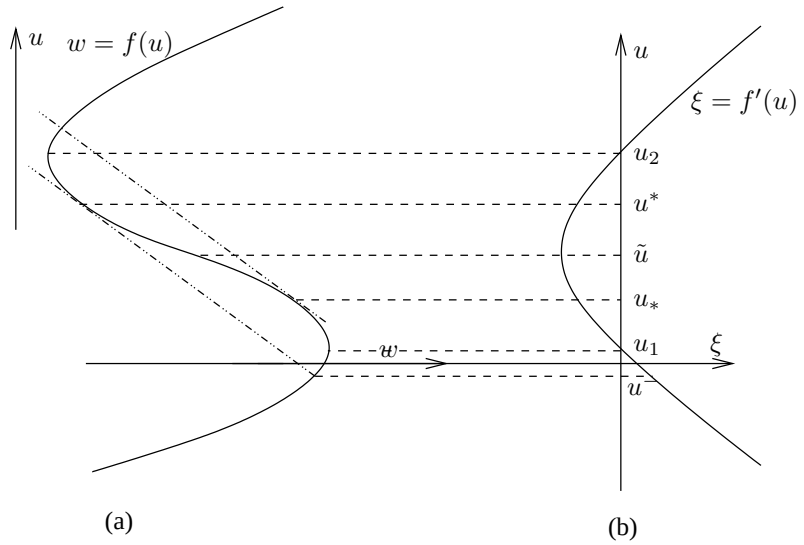


FIG. 3.1. *The graphs of* $w = f(u)$ *and* $\xi = f'(u)$ *when* $f'(\pm\infty) = +\infty$.

consider the case that $f'(\tilde{u}) < 0$. Then there are two critical points $u_1 < u_2$ such that $f'(u_1) = f'(u_2) = 0$ since $f'(\tilde{u}) < 0$ and $f'(\pm\infty) = +\infty$. For definiteness, we assume $u_1 > 0$. The graphs of $w = f(u)$ in the (w, u) plane and $\xi = f'(u)$ in the (ξ, u) plane are shown in Figures 3.1(a) and (b).

Let $u^- < 0$ be a given state. Since our attention in this paper is paid on nonconvex cases, we assume that $u^- < u_1$. If $f(u^-) \le f(u_2)$, then we have a unique admissible noncombustion solution consisting of a forward wave (a shock or a compound wave—a sonic shock plus a rarefaction wave),

$$(u(\xi), z(\xi)) = \begin{cases} (u^-, 1), & \xi < 0, \\ (u^-, 0), & \xi \in (0, \bar{\xi}), \\ (u^+, 0), & \xi > \bar{\xi}, \end{cases}$$

where $\bar{\xi} = \frac{f(v)-f(u^-)}{v-u^-} = f'(v)$. Therefore, it is sufficient to consider the case that $f(u^-) > f(u_2)$.

First, we fix the notation $\bar{u}$, u^*, u_*, $\bar{q}$, q_*, a_0, a_1, and a_2. Let $\bar{u}$, u^*, u_* be so defined that $u_1 < \bar{u} < u_2$, $u^* > u_*$, $f(\bar{u}) = f(u^-)$, and

$$f'(u^*) = \frac{f(u^*) - f(u^-)}{u^* - u^-} = f'(u_*). \tag{3.1}$$

Let $\bar{q}$, q_* be such that

$$\bar{q} = \bar{u} - u^-, \qquad \frac{f(u_*) - f(u^-)}{u_* - (u^- + q_*)} = f'(u_*). \tag{3.2}$$

Then we have $u^* > \bar{u}$ and $\bar{q} < q_*$. Note that q_* is the distance between the tangential lines of $f(u)$ at $(u^*, f(u^*))$ and at $(u_*, f(u_*))$ vertically. We restrict ourselves to dealing with the case where $u_* > \bar{u}$ and $0 < q < q_*$ since the other cases can be treated similarly.

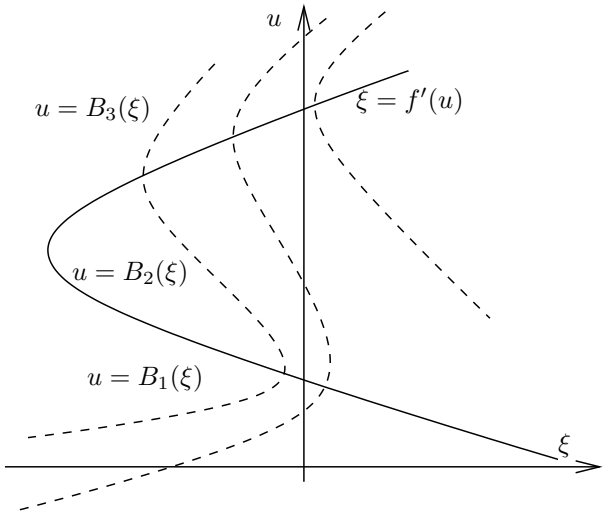


FIG. 3.2. *The branches of integral curve of* (3.3).

Draw a straight line $w - f(u^-) = f'(u^*)(u - (u^- + q))$ in the (u, w) plane, which intersects $w = f(u)$ at three points $(a_i, f(a_i)), i = 0, 1, 2$, $a_0 < a_1 < a_2$. Then we have the relation

$$u_2 > u^* > a_1 > u_* > a_0 > u_-.$$

We also denote $\eta^* = f'(u^*)$.

Fix u^- and q. Then the structure of solution will strongly depend on the value of u^+. Therefore we discuss the issue by the following three cases.

Case 3.1.1. $u^+ \in (a_1, +\infty)$.

Case 3.1.2. $u^+ \in (\bar{u}, a_1]$.

Case 3.1.3. $u^+ \in (0, \bar{u})$.

Before proceeding to discuss these cases, some lemmas are given in the following.

LEMMA 3.1. *All integral curves of the ordinary differential equation*

$$(f'(u) - \xi)\frac{du}{d\xi} = qk\left(\frac{\xi}{\eta}\right)^k \tag{3.3}$$

intersecting with $\xi = f'(u)$ *in the* (ξ, u) *plane are shown in Figure* 3.2. *Each of the integral curves consists of three branches:* $u = B_i(\xi)$, $i = 1, 2, 3$, *having the property that* $B_3(\xi) > R_2(\xi) > B_2(\xi) > R_1(\xi) > B_1(\xi)$.

Proof. Denote the integral curve of (3.3) by $u = B(\xi)$. Let (ξ_0, u_0) be the intersection point of $u = B(\xi)$ and $u = R_2(\xi)$. Then (ξ_0, u_0) is the singularity point of (3.3).

Denote by $u = B_3(\xi)$ the branch that lies in the upper side of $u = R_2(\xi)$ in the neighborhood of (ξ_0, u_0), i.e., $B_3(\xi) > R_2(\xi)$. Then it suffices to show that $B_3(\xi)$ will not intersect with $R_2(\xi)$ for the finite reaction rate $k < \infty$ as $\xi > \xi_0$. Indeed, using

(3.3), we obtain

$$\frac{d}{d\xi}(B_3(\xi)-R_2(\xi))=\frac{kq(\frac{\xi}{\eta})^k}{f'(B_3(\xi))-\xi}-\frac{1}{f''(R_2(\xi))}$$
$$=\frac{kq(\frac{\xi}{\eta})^k f''(R_2(\xi))-(f'(B_3(\xi))-\xi)}{(f'(B_3(\xi))-\xi)f''(R_2(\xi))}.$$

Recall that $u=R_2(\xi)$ is the branch associated with the convex part of $f(u)$. Then $f''(R_2(\xi))>0$. Observe that once $u=B_3(\xi)$ is close to $u=R_2(\xi)$, $f'(B_3(\xi))-\xi$ becomes small. Therefore, at this moment, the numerator is positive, which forces $u=B_3(\xi)$ to leave away $u=B_2(\xi)$ for the upper side of $R_2(\xi)$.

Similarly, we can prove the lemma for $u=B_2(\xi)$ and $u=B_1(\xi)$. □

The solution $u=B_i(\xi)$ $(i=1,2,3)$ of (3.3) depends on the parameters η and k. When there is no risk of confusion, we suppress the dependence of B_i on η and k. Otherwise, we denote $u=B_i(\xi;\eta,k)$.

LEMMA 3.2. *Let $u=B_i(\xi)$, $i=2,3$, be the branches of integral curve of* (3.3) *through the point (η^*,u^*). Then*

(a) $\lim_{k\to+\infty}B_3(\xi)=\max\{a_2,R_2(\xi)\}$ *for all* $\xi\in(\eta^*,0]$;

(b) $\lim_{k\to+\infty}B_2(\xi)=\max\{a_1,R_1(\xi)\}$ *for all* $\xi\in(\eta^*,0]$.

Proof. We prove in two steps part (a) only. Part (b) can be treated in the same way. Note that $B_3(\xi)$ depends on k.

(i) The first step is to prove

$$\underline{\lim_{k\to+\infty}}B_3(\xi)\geq\max\{a_2,R_2(\xi)\}.$$

We write $\underline{\lim_{k\to+\infty}}=:\underline{\lim}$ for short. Since $B_3(\xi)\geq R_2(\xi)$, it suffices to prove $\underline{\lim}B_3(\xi)\geq a_2$. Assume to the contrary that this inequality is not true. Then there exists $\xi_0\in(\eta^*,0]$ such that $\underline{\lim}B_3(\xi_0)<a_2$. Setting $\eta=\eta^*$, $B_3(\eta^*)=u^*$ in (3.3), and integrating (3.3), from η^* to ξ_0, we get

(3.4)
$$f(B_3(\xi_0))-f(u^*)-\xi_0B_3(\xi_0)+\eta^*u^*+\int_{\eta^*}^{\xi_0}B_3(\xi)d\xi=\frac{qk\eta^*}{k+1}\left[\left(\frac{\xi_0}{\eta^*}\right)^{k+1}-1\right].$$

Since $B_3(\xi)$ is increasing, we have

$$\int_{\eta^*}^{\xi_0}B_3(\xi)d\xi<B_3(\xi_0)(\xi_0-\eta^*).$$

Letting $k\to+\infty$ in (3.4) gives

$$f(\underline{B})-f(u^*)-\eta^*(\underline{B}-u^*-q)\geq 0, \tag{3.5}$$

where $\underline{B}=\underline{\lim}B_3(\xi_0)$. Since a_2 is so defined that

$$f(a_2)-f(u^-)=f'(u^*)(a_2-(u^-+q)),$$

we get from (3.5)

$$f(a_2)-f(\underline{B})-\eta^*(a_2-\underline{B})\leq 0, \tag{3.6}$$

where we use the definition of u^*, $f'(u^*)(u^*-u^-)=f(u^*)-f(u^-)<0$.

On the other hand, it follows from $u^*\le \underline{B}=\underline{\lim}B_3(\xi_0)<a_2$ and the convexity of $f(u)$ that

$$\frac{f(a_2)-f(\underline{B})}{a_2-\underline{B}}>\eta^*, \tag{3.7}$$

which contradicts (3.6). Hence $\underline{\lim}B_3(\xi)\ge a_2$ for all $\xi\in(\eta^*,0]$.

(ii) The second step is to verify

$$\overline{\lim_{k\to+\infty}}B_3(\xi)\le\max\{a_2,R_2(\xi)\}.$$

For any $\xi_0\in(\eta^*,0]$, we get from (3.4) that

$$\begin{aligned}f(B_3(\xi_0))&\le f(u^*)+\xi_0B_3(\xi_0)-\int_{\eta^*}^{\xi_0}B_3(\xi)d\xi-\eta^*(u^*+q)\\&\le f(u^*)+\xi_0B_3(\xi_0)-u^*(\xi_0-\eta^*)-\eta^*(u^*+q)\\&<f(u^*)-\eta^*q,\end{aligned}$$

since $\xi_0<0$ and $B_3(\xi_0)>u^*$. This implies that $\{B_3(\xi_0)\}$ is bounded uniformly. Hence $\overline{\lim}B_3(\xi_0):=\overline{\lim_{k_i\to+\infty}}B_3(\xi_0)$ exists for any $\xi_0\in(\eta^*,0]$, denoted by $\overline{B}(\xi_0)$. Choose a subsequence $\{k_i\}$ from $\{k\}$ such that

$$\lim_{k_i\to+\infty}B_3(\xi_0)=\overline{B}(\xi_0).$$

Using Fatou's lemma, we get

$$\begin{aligned}\overline{\lim_{k_i\to+\infty}}\int_{\eta^*}^{\xi_0}B_3(\xi)d\xi&\ge\varlimsup_{k_i\to+\infty}\int_{\eta^*}^{\xi_0}B_3(\xi)d\xi\ge\int_{\eta^*}^{\xi_0}\varlimsup_{k_i\to+\infty}B_3(\xi)d\xi\\&\ge\int_{\eta^*}^{\xi_0}\max(a_2,R_2(\xi))d\xi.\end{aligned}\tag{3.8}$$

Setting $k_i\to+\infty$ in (3.4) and noting (3.8), we arrive at

$$f(\overline{B}(\xi_0))-f(u^*)-\xi_0\overline{B}(\xi_0)+\eta^*u^*+\int_{\eta^*}^{\xi_0}\max\{a_2,R_2(\xi)\}d\xi\le-q\eta^*. \tag{3.9}$$

By the definition of a_2, we have $a_2>u^*$. Let $\bar\xi\in(\eta^*,\xi_0]$ such that $R_2(\bar\xi)=a_2$. (If $a_2>R_2(\xi)$ for all $\xi\in(\eta^*,\xi_0]$, then we take $\bar\xi=\xi_0$.) Then

$$\int_{\eta^*}^{\bar\xi}a_2d\xi=(\bar\xi-\eta^*)a_2$$

and

$$\int_{\bar\xi}^{\xi_0}R_2(\xi)d\xi=f(a_2)-\bar\xi a_2-(f(R_2(\xi_0))-\xi_0R_2(\xi_0)).$$

Note that

$$f(u^*)-f(u^-)=\eta^*(u^*-u^-),\qquad f(a_2)=f(u^-)+\eta^*(a_2-(u^-+q)).$$

Then it is calculated that

$$(\bar{\xi}-\eta^*)a_2 = f(u^*) - f(a_2) + \bar{\xi}a_2 - \eta^*(u^*+q).$$

Consequently, we obtain

$$\begin{aligned}(3.10)\qquad &\int_{\eta^*}^{\xi_0} \max\{a_2, R_2(\xi)\}d\xi \\ &= f(u^*) - f(\max\{a_2, R_2(\xi_0)\}) + \xi_0 \max\{a_2, R_2(\xi_0)\} - \eta^*(u^*+q),\end{aligned}$$

which, combining with (3.9), provides

$$(3.11)\qquad f(\overline{B}(\xi_0)) - f(\max\{a_2, R_2(\xi_0)\}) - \xi_0(\overline{B}(\xi_0) - \max\{a_2, R_2(\xi_0)\}) \le 0.$$

Then there is $\theta \in (\max\{a_2, R_2(\xi_0)\}, \overline{B}(\xi_0))$ such that

$$f(\overline{B}(\xi_0)) - f(\max\{a_2, R_2(\xi_0)\}) = f'(\theta)(\overline{B}(\xi_0) - \max\{a_2, R_2(\xi_0)\}).$$

Note that $f''(u) > 0$ for all $u > \tilde{u}$ and $\overline{B}(\xi_0) > \theta > \max\{a_2, R_2(\xi_0)\} > \tilde{u}$. We have

$$f'(\theta) - \xi_0 = f'(\theta) - f'(R_2(\xi_0)) > 0.$$

Hence, (3.11) implies $\overline{B}(\xi_0) - \max\{a_2, R_2(\xi_0)\} < 0$. □

LEMMA 3.3. *Let $u = B_2(\xi)$ be the integral curve of* (3.3) *through* $(0, u^+)$, *where* $u_1 < u^+ < u_2$. *Then*

$$(3.12)\qquad \lim_{k\to+\infty} B_2(\xi) = \max\{u^+, R_1(\xi)\} \quad \textit{for all } \xi \in (f'(u^+), 0].$$

Proof. By Lemma 3.1, we have for a finite reaction rate k,

$$B_2(\xi) > u^+, \quad B_2(\xi) > R_1(\xi) \quad \text{for all } \xi \in (f'(u^+), 0],$$

and $B_2(\xi)$ is decreasing. Note that $u^+ > R_1(\xi)$ for all $\xi \in (f'(u^+), 0)$. Hence it suffices to prove that

$$(3.13)\qquad \lim_{r\to\infty} B_2(\xi) = u^+.$$

Assume to the contrary that there exists $\xi_0 \in (f'(u^+), 0)$ such that $\underline{\lim}_{k\to\infty} B_2(\xi_0) =: v > u^+$. Then we integrate (3.3) from ξ_0 to 0 and obtain

$$f(u^+) - f(B_2(\xi_0)) + \xi_0 B_2(\xi_0) + \int_{\xi_0}^0 B_2(\xi)d\xi = \frac{kq\xi_0}{k+1}\left(\frac{\xi_0}{\eta}\right)^k.$$

Since

$$\int_{\xi_0}^0 B_2(\xi)d\xi > -\xi_0 u^+,$$

we obtain that as k goes to infinity,

$$f(u^+) - f(v) + \xi_0(v - u^+) < 0,$$

i.e.,

$$\frac{f(u^+)-f(v)}{u^+-v}>\xi_0.$$

Therefore, there exists $\theta\in(u^+,v)$ such that

$$f'(\theta)=\frac{f(u^+)-f(v)}{u^+-v}>\xi_0. \tag{3.14}$$

(a) If $\theta\in(u_1,\tilde{u})$, then since $f''(u)<0$ and $v>\theta>u^+$, we have $f'(v)<f'(\theta)<f'(u^+)$. This contradicts the fact that $f'(\theta)>\xi_0>f'(u^+)$.

(b) If $\theta\in[\tilde{u},u_2)$, then $f'(v)>f'(\theta)$ since $f''(u)>0$ for all $u\in(\tilde{u},u_2)$. Note that $\xi_0>f'(v)$. We obtain the contradiction to (3.14). □

In the remainder of this section, we will suppress the subscript of B when doing so causes no confusion.

LEMMA 3.4. *Let $\eta\in(\eta^*,0)$ and $B(0;\eta,k)=u^+\in(\bar{u},a_1)$. Then there exists a constant $k_0>0$ such that whenever $k>k_0$,*

$$B(\xi;\eta,k)<u^* \quad \textit{uniformly for all } \ \xi\in[\eta,0],\ \eta\in(\eta^*,0). \tag{3.15}$$

Proof. If (3.15) fails, then for any $n>0$ there exist $k_n>n$, $\eta_n\in(\eta^*,0)$, and $\xi_n\in[\eta_n,0]$ such that $B(\xi;\eta_n,k_n)>u^*$ for all $\xi\in(\xi_n,0]$ and $B(\xi_n;\eta_n,k_n)=u^*$.

Since $\eta_n\in(\eta^*,0)$, we choose a convergent subsequence from $\{\eta_n\}$, still denoted by $\{\eta_n\}$. Set $\overline{\eta}=\lim\eta_n$, $\overline{\xi}=\lim\xi_n$. Substituting η, k, and u by η_n, k_n, and $B(\xi;\eta_n,k_n)$ in (3.3), respectively, and then integrating it from ξ_n to 0, we obtain

$$f(u^+)-f(u^*)+\xi_n u^*+\int_{\xi_n}^{0}B(\xi;\eta_n,k_n)d\xi=-q\xi_n\frac{k_n}{k_n+1}\left(\frac{\xi_n}{\eta_n}\right)^{k_n}. \tag{3.16}$$

Letting $n\to+\infty$ and noting $B(\xi;\eta_n,k_n)\geq u^+$, we have from (3.16) that

$$\frac{f(u^*)-f(u^+)}{u^*+q-u^+}\geq\overline{\xi}. \tag{3.17}$$

On the other hand, when $u^+\in(\bar{u},a_1)$,

$$\frac{f(u^*)-f(u^+)}{u^*+q-u^+}<\eta^*. \tag{3.18}$$

Then $\eta^*>\overline{\xi}$, which contradicts $\overline{\xi}\geq\overline{\eta}\geq\eta^*$. Thus, we complete the proof. □

LEMMA 3.5. *Let $B(0;\eta_1,k)=B(0;\eta_2,k)=u^+\in(\bar{u},a_1)$. If $\eta_1\geq\eta_2>\eta^*$, then, for sufficiently large k,*

$$B(\xi;\eta_1,k)>B(\xi;\eta_2,k) \textit{ for all } \xi\in[\eta_1,0]. \tag{3.19}$$

Proof. Letting $w(\xi;\eta,k)=\frac{\partial B(\xi;\eta,k)}{\partial\eta}$, and differentiating (3.3) with respect to ξ, we obtain

$$\begin{cases}\dfrac{dw}{d\xi}+\dfrac{f''(B)qk(\frac{\xi}{\eta})^k}{(f'(B)-\xi)^2}w=\dfrac{-qk(\frac{\xi}{\eta})^k}{\eta(f'(B)-\xi)}, \quad \eta<\xi<0,\\ w(0;\eta,k)=0.\end{cases}$$

From Lemma 3.1, it follows that $f'(B(\xi;\eta,k))-\xi<0$. Then the fact that the right-hand side of the above equation is negative implies $w>0$ for all $\xi\in(\eta,0)$. Hence, $B(\xi;\eta,k)$ increases in η. Thus we obtain (3.19). ∎

LEMMA 3.6. *Let $B(0;\eta,k)=u^+\in(\bar{u},a_1)$. Then, for sufficiently large k, there exists $\eta_k\in(\eta^*,0)$ such that*

$$\eta_k=\frac{f(B(\eta_k;\eta_k,k))-f(u^-)}{B(\eta_k;\eta_k,k)-u^-}. \tag{3.20}$$

Proof. Let $H_k(\eta)=f(B(\eta;\eta,k))-f(u^-)-\eta(B(\eta;\eta,k)-u^-)$. By Lemma 3.1, we have $u^+<B(\eta^*;\eta^*,k)<u^*$. Then

$$H_k(\eta^*)=f(B(\eta^*;\eta^*,k))-f(u^-)-\eta^*(B(\eta^*;\eta^*,k)-u^-)>0$$

and

$$H_k(0)=f(u^+)-f(u^-)-0(u^+-u^-)<0,$$

which implies the existence of η_k satisfying (3.20). ∎

Now we turn to consider the admissible solution of (2.2) and (2.4) corresponding to Cases 3.1.1–3.1.3. In the following arguments, we need to notice the dependence of solutions $(u(\xi),z(\xi))$ of (2.2) and (2.4) on the reaction rate k.

Case 3.1.1. $u^+\in(a_1,+\infty)$.

THEOREM 3.7. *For Case* 3.1.1, *the Riemann problem* (2.2) *and* (2.4) *has a unique admissible solution $(u(\xi),z(\xi))$,*

$$(u(\xi),z(\xi))=\begin{cases}(u^-,1), & \xi<\eta^*,\\ (B(\xi;\eta^*,k),(\frac{\xi}{\eta^*})^k), & \xi\in[\eta^*,0],\\ (C(\xi),0), & \xi>0,\end{cases} \tag{3.21}$$

where $C(\xi)$ is the solution of the following boundary value problem:

$$\begin{cases}(f'(C)-\xi)\frac{dC}{d\xi}=0, C(\xi)\geq 0, \quad \xi\in(0,+\infty),\\ C(0)=B(0;\eta^*,k),\\ C(+\infty)=u^+.\end{cases} \tag{3.22}$$

The infinite reaction rate limit is

$$\lim_{k\to+\infty}(u(\xi),z(\xi))=\begin{cases}(u^-,1), & \xi<\eta^*,\\ (U(\xi),0), & \xi\geq\eta^*,\end{cases} \tag{3.23}$$

where $U(\xi)$ is the solution of

$$\begin{cases}(f'(U)-\xi)\frac{dU}{d\xi}=0, \eta^*<\xi<+\infty,\\ U(\eta^*)=a_2, \quad U(+\infty)=u^+.\end{cases} \tag{3.24}$$

Note that here $\eta^*=\frac{f(a_2)-f(u^-)}{a_2-(u^-+q)}<f'(a_2)$.

Proof. Note that $B_3(0;\eta^*,k)>u_2$ and $\lim_{k\to\infty}B_3(\xi;\eta^*,k)=\max\{R_2(\xi),a_2\}$ for $\xi\in[\eta^*,0]$. Also note that

$$\frac{f(u_2)-f(u)}{u_2-u}\geq 0 \quad \text{for } u\in[u_2,\infty) \tag{3.25}$$

and

$$\frac{f(u_2)-f(u)}{u_2-u}<0 \quad \text{for } u\in(a_1,u_2). \tag{3.26}$$

Then there exists $v\in(a_1,u_2)$ such that for the sufficiently large k, $f(B_3(0;\eta^*,k))=f(v)$. Thus we have two subcases:

(i) $u^+\in[v,\infty)$. Then

$$\frac{f(B_3(0;\eta^*,k))-f(u)}{B_3(0;\eta^*,k))-u}\geq 0 \quad \text{for all } u\in[v,\infty). \tag{3.27}$$

Hence the boundary problem (3.22) has the unique solution $C(\xi)$ for all $\xi>0$ with $C(0+0)=B_3(0;\eta^*,k)$. Thus the solution $(u(\xi),z(\xi))$ can be expressed as in (3.21), where $u(\xi)=B_3(\xi;\eta^*,k)$ as $\xi\in[\eta^*,0)$. This is obviously the admissible solution of (2.2) and (2.4), as shown in Figure 3.3(a).

Next we prove the uniqueness. Let $(u(\xi),z(\xi))$ be the solution of (2.2) and (2.4) for this case. According to Lemma 2.1, $z(\xi)$ has the structure (2.8). When $\xi<\eta$, u satisfies (2.10), of which the unique solution is $u=u^-$. When $\xi\geq 0$, u satisfies (2.12). When $u^+>u_2$, there exists a unique solution of (2.12) for a fixed $u(0+0)$ if $u(0+0)>u_2$. So, we only need to prove $\eta=\eta^*$ and $u(\xi)=B(\xi;\eta^*,k)$, as $\xi\in(\eta^*,0)$.

Indeed, we know from the Rankine–Hugoniot jump condition (2.6) that $\eta\geq\eta^*$. If $\eta>\eta^*$, then the Rankine–Hugoniot jump condition (2.6) and the Oleinik-type entropy condition (2.7) imply $u(\eta+0)<u^*$. Therefore the solution of (2.11) must be continuous and decreasing. Thus, $u(0-0)<u^*$. However, the discontinuity, which is the jump at $\xi=0$ from $u(0-0)<u^*$ to $u(0+0)>u_2$, does not satisfy the Oleinik-type entropy condition (2.7). Therefore, $\eta=\eta^*$ and $u(\eta^*+0)=u^*$. Note from (2.9) that $u(\xi)=B_3(\xi;\eta^*,k)$ lies in the right-hand side of η^*. We conclude that $u(\xi)$ is continuous in $(\eta^*,0)$. Otherwise, suppose ξ_0 to be a discontinuity point of $u(\xi)$. Then, by the Oleinik-type entropy condition (2.7) (observe Figure 3.1), it is easy to see that $u(\xi_0+0)<u^*$, which, combined with the entropy condition, implies that $u(\xi)$ is decreasing and continuous in $(\eta^*,0)$. Consequently, $u(0-0)<u^*$. However, the discontinuity at $\xi=0$ does not satisfies the Oleinik-type entropy condition (2.7). This is a contradiction.

By Lemma 3.2, we have

$$\lim_{k\to\infty} B_3(\xi;\eta^*,k)=\max\{a_2,R_2(\xi)\},\quad \xi\in(\eta^*,0),$$

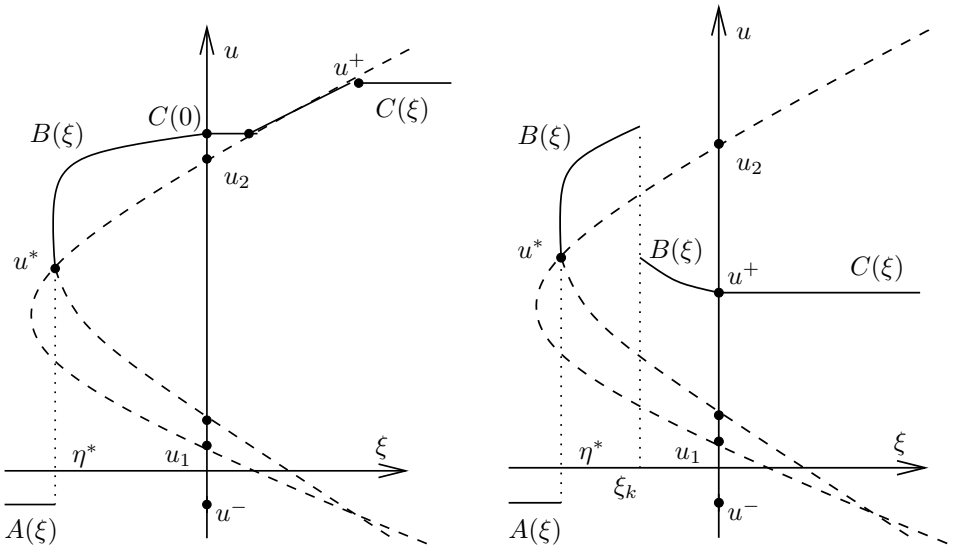
which is the same as the solution of (3.24) for $\xi\in(\eta^*,0)$.

(ii) $u^+\in[a_1,v)$. Note that

$$\frac{f(B_3(0;\eta^*,k))-f(u)}{B_3(0;\eta^*,k))-u}<0 \quad \text{for all } u\in[a_1,v). \tag{3.28}$$

Then we cannot find a solution to (3.22) such that $C(0+0)=B_3(0;\eta^*,k)$. Instead, we construct a solution in the form of (3.21), in which $B(\xi;\eta^*,k)$, at this moment, becomes

$$B(\xi;\eta^*,k)=\begin{cases}B_3(\xi;\eta^*,k), & \xi\in[\eta^*,\xi_k),\\ B_2(\xi;\eta^*,k), & \xi\in[\xi_k,0),\end{cases} \tag{3.29}$$



(a) The solution when $u^+ \in (v, +\infty)$.

(b) The solution when $u^+ \in (a_1, v)$.

FIG. 3.3. *The structure of the entropy solution to the ZND model when* $u^+ \in (a_1, \infty)$.

where $B_3(\xi;\eta^*,k) =: B_3(\xi)$ and $B_2(\xi;\eta^*,k) =: B_2(\xi)$ are defined in Lemmas 3.2 and 3.3, respectively, and $\xi_k \in [\eta^*, 0)$ is to be determined. Next we show that this is indeed the solution of (2.2) and (2.4) for a suitable ξ_k.

In fact, set $\eta = \eta^*$ in (3.3). Then we integrate (3.3) from η^* to ξ_k and from ξ_k to 0, respectively, to obtain

(3.30)
$$f(B_3(\xi_k)) - f(u^*) - \xi_k B_3(\xi_k) + \eta^* u^* + \int_{\eta^*}^{\xi_k} B_3(\xi)d\xi = \frac{kq\eta^*}{k+1}\left(\left(\frac{\xi_k}{\eta^*}\right)^{k+1} - 1\right)$$

and

$$f(u^+) - f(B_2(\xi_k)) + \xi_k B_2(\xi_k) + \int_{\xi_k}^{0} B_2(\xi)d\xi = -\frac{kq\eta^*}{k+1}\left(\frac{\xi}{\eta^*}\right)^{k+1}. \tag{3.31}$$

We add these two identities to get

$$\begin{aligned} &f(B_2(\xi_k)) - f(B_2(\xi_k)) - \xi_k(B_3(\xi_k) - B_2(\xi_k)) \\ &+ f(u^+) - f(u^*) + \eta^* u^* + \frac{kq\eta^*}{k+1} + \int_{\eta^*}^{\xi_k} B_3(\xi)d\xi + \int_{\xi_k}^{0} B_2(\xi)d\xi = 0. \end{aligned} \tag{3.32}$$

Set

$$H(\xi) = f(B_3(\xi)) - f(B_2(\xi)) - \xi(B_3(\xi) - B_2(\xi)). \tag{3.33}$$

Then

$$\frac{dH}{d\xi} = B_2(\xi) - B_3(\xi) < 0. \tag{3.34}$$

Note that

$$\frac{f(a_2)-f(u^+)}{a_2-u^+}>\eta^*. \tag{3.35}$$

Hence, in light of Lemmas 3.2 and 3.3, for the sufficiently large k, there exists $\eta^*<\bar{\xi}<0$ such that $H(\bar{\xi})>0$. Since $f(B_3(0))=f(v)<f(u^+)$, $H(0)<0$. Therefore, there exists a unique $\xi_k\in(\bar{\xi},0)$ such that $H(\xi_k)=0$. Take

$$\xi_k=\frac{f(B_3(\xi_k))-f(B_2(\xi_k))}{B_3(\xi_k)-B_2(\xi_k)}. \tag{3.36}$$

Then the solution $u=u(\xi)$ has a discontinuity at $\xi=\xi_k$, which satisfies the Oleinik entropy condition (2.7) since $B_3(\xi_k)>B_2(\xi_k)$ and $B_3(\xi_k)>u^*$.

In light of (3.32), the solution $(u(\xi),z(\xi))$ constructed above satisfies (2.2) and (2.4) in the sense of distributions and therefore is admissible. We display this solution in Figure 3.3(b).

The uniqueness can be proved similarly to that in (i).

By Lemmas 3.2 and 3.3, the limit of this solution as the reaction rate k goes to infinity is expressed in (3.23). □

Case 3.1.2. $u^+\in(\bar{u},a_1]$.

THEOREM 3.8. *When k is sufficiently large, then*

(1) *there exists $\eta_k\in(\eta^*,0)$ such that the Riemann problem for Case* 3.1.2 *has the unique admissible solution*

$$(u(\xi),z(\xi))=\begin{cases}(u^-,1), & \xi\in(-\infty,\eta_k),\\ (B(\xi;\eta_k,k),(\frac{\xi}{\eta_k})^k), & \xi\in(\eta_k,0),\\ (u^+,0), & \xi\in(0,+\infty),\end{cases} \tag{3.37}$$

where $B(\xi;\eta_k,k)$ satisfies (3.3) *with $B(0;\eta_k,k)=u^+\in(\bar{u},a_1]$ and*

$$\eta_k=\frac{f(B(\eta_k;\eta_k,k))-f(u^-)}{B(\eta_k;\eta_k,k)-u^-};$$

(2) *there holds*

$$\lim_{k\to+\infty}(u(\xi),z(\xi))=\begin{cases}(u^-,1), & \xi\in(-\infty,\eta_0),\\ (U(\xi),0), & \xi\in(\eta_0,+\infty),\end{cases} \tag{3.38}$$

where $U(\xi)=\max\{R_1(\xi),u^+\}$ and η_0 satisfies

$$\eta_0=\frac{f(U(\eta_0))-f(u^-)}{U(\eta_0)-u^--q}\geq f'(U(\eta_0)). \tag{3.39}$$

Proof. (1) If $(u(\xi),z(\xi))$ is the solution of (2.2) and (2.4), then $u(\xi)$ satisfies (2.10) and (2.12) when $\xi\in(-\infty,\eta)$ and $\xi\in(0,+\infty)$, respectively. By the Oleinik-type entropy condition (2.7) and under the assumption that $u^-<a_0$ and $u^+\in(\bar{u},a_1]$, we have $u(\xi)=u^-$ for $\xi\in(-\infty,\eta)$ and $u(\xi)=u^+$ for $\xi\in(0,+\infty)$. By Lemma 3.6, there exists $\eta_k\in(\eta^*,0)$ such that

$$\eta_k=\frac{f(B(\eta_k;\eta_k,k))-f(u^-)}{B(\eta_k;\eta_k,k)-u^-}.$$

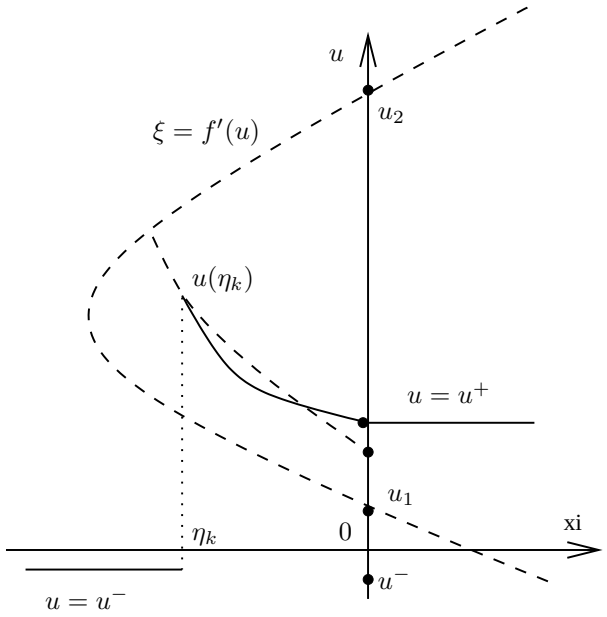


FIG. 3.4. *The structure of entropy solution to the ZND model when* $u^+ \in (\bar{u}, a_1)$.

It is evident that the entropy condition (2.7) is satisfied at the jump $\xi = \eta_k$. Thus, the admissible solution exists (see Figure 3.4). Arguments similar to those in Theorem 3.7 show that such a solution is unique.

(2) Integrating (3.3) from η_k to 0, we have

$$f(u^+) - f(B(\eta_k;\eta_k,k)) + \eta_k B(\eta_k;\eta_k,k) + \int_{\eta_k}^{0} B(\xi;\eta_k,k)d\xi + q\eta_k \frac{k}{k+1} = 0. \tag{3.40}$$

Substituting $H(\eta_k) = 0$ (where $H(\eta_k)$ is defined in Lemma 3.6) into (3.40) gives

$$f(u^+) - f(u^-) + \eta_k u^- + q\eta_k \frac{k}{k+1} + \int_{\eta_k}^{0} B(\xi;\eta_k,k)d\xi = 0. \tag{3.41}$$

Let $\bar{\eta}_0 = \overline{\lim_{k\to+\infty}}\eta_k$. Then, in light of Lemma 3.3, we get from (3.41)

$$f(u^+) - f(u^-) + \bar{\eta}_0 u^- + q\bar{\eta}_0 + \int_{\bar{\eta}_0}^{0} U(\xi)d\xi = 0, \tag{3.42}$$

where $U(\xi) = \max\{u^+, R_1(\xi)\}$. Therefore, we have

$$\int_{\bar{\eta}_0}^{0} U(\xi)d\xi = \begin{cases} -\bar{\eta}_0 u^+, & \bar{\eta}_0 > f'(u^+), \\ -\bar{\eta}_0 U(\bar{\eta}_0) - f(u^+) + f(U(\bar{\eta}_0)), & \bar{\eta}_0 \le f'(u^+). \end{cases} \tag{3.43}$$

Equations (3.42) and (3.43) imply that $\bar{\eta}_0$ satisfies (3.39).

Analogously, it is proved that $\underline{\eta}_0 = \lim_{\overline{k\to+\infty}}\eta_k$ satisfies (3.39). Therefore, $\eta_0 = \lim_{\overline{k\to+\infty}}\eta_k = \bar{\eta}_0 = \underline{\eta}_0$ and satisfies (3.39).

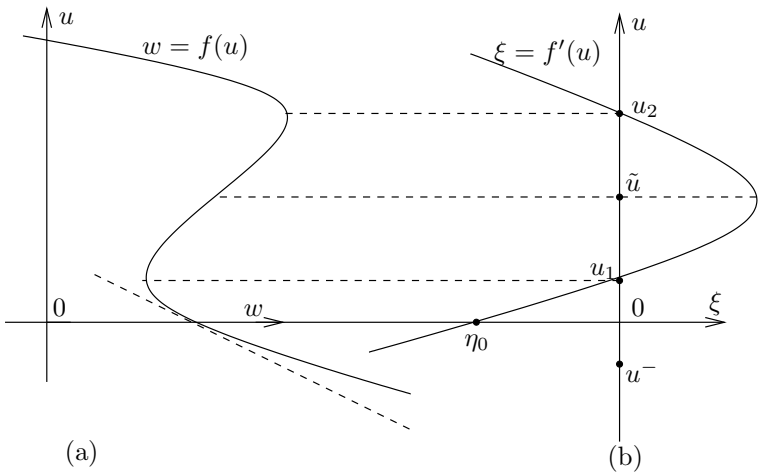


FIG. 3.5. *The graphs of* $w = f(u)$ *and* $\xi = f'(u)$ *when of* $f'(\pm\infty) = -\infty$.

For any given $\xi_0 \in (\eta_0, 0)$, choose a small $\epsilon > 0$ so that $\eta_0 + \epsilon < \xi_0$. Then we have $\eta_k < \eta_0 + \epsilon < \xi_0$ for sufficiently large k. From Lemma 3.6, we get

$$U(\xi_0) < B(\xi_0; \eta_k, k) < B(\xi_0; \eta_0 + \epsilon, k).$$

Thus, letting $k \to +\infty$, we obtain

$$\lim_{k\to\infty} B(\xi_0; \eta_k, k) = U(\xi_0).$$

Due to the arbitrariness of ξ_0, $\lim_{k\to\infty} B(\xi; \eta_k, k) = U(\xi)$ for all $\xi \in (\eta_0, 0)$. □

Case 3.1.3. $\boldsymbol{u^+ \in (0, \bar{u}]}$**.** For this case, we have the following theorem.

THEOREM 3.9. *For Case* 3.1.3, *the unique admissible solution of* (2.2) *and* (2.4) *is the noncombustion solution*

$$(u(\xi), z(\xi)) = \begin{cases} (u^-, 1), & \xi < \frac{f(u^-)-f(u^+)}{u^- - u^+}, \\ (u^+, 0), & \xi > \frac{f(u^-)-f(u^+)}{u^- - u^+}. \end{cases} \tag{3.44}$$

3.2. The solution of (2.2) and (2.4) when $\boldsymbol{f(u)}$**satisfies** $\boldsymbol{(A_2)}$**.** Parallel to section 3.1, we solve (2.2) and (2.4) as $f(u)$ has just one inflection point and the slope at infinity is negative infinity. The results are stated in Theorems 3.12 and 3.13. We omit some proofs because they are similar to those in section 3.1.

Let $\tilde{u}$ be the inflection point of $f(u)$. When $f'(\tilde{u}) \le 0$, there is no combustion solution (cf. [13]). Therefore, we only need to consider $f'(\tilde{u}) > 0$. Let u_1, u_2, and u_3 be such that $u_3 > u_2 > u_1 > 0$ and $f'(u_1) = f'(u_2) = 0$. Recall that $u = 0$ is assumed to be the ignition point and therefore plays an important role in the current context. We set $\hat{q}$ to satisfy $f'(0) = \frac{f(u_2)-f(0)}{u_2-\hat{q}}$. The value of the binding energy q is distinguished into two classes: $0 < q < \hat{q}$ and $q \ge \hat{q}$. We restrict our study to the solution to the case that $q \in (0, \hat{q})$. The case that $q \ge \hat{q}$ can be treated similarly.

Draw the straight line $w - f(0) = f'(0)(u - q)$ in the (u, w) plane. Then it intersects $w = f(u)$ at three points $(a_i, f(a_i))$, $i = 1, 2, 3$, where $a_2 > a_1 > a_0$. It is evident that $a_2 > u_2 > a_1 > 0 > a_0$.

For the fixed $q \in (0, \hat{q})$ and $u^- \in (-\infty, 0)$, the structure of the solution of the Riemann problem (2.2) and (2.4) will depend on the value of u^+. We proceed our discussion through two cases.

Case 3.2.1. $u^+ \in (0, a_2)$.

Case 3.2.2. $u^+ \in [a_2, \infty)$.

Case 3.2.1. $u^+ \in (0, a_2)$. The main result is stated in Theorem 3.12 by following Lemmas 3.10 and 3.11.

LEMMA 3.10. *Let $B(\xi; \eta_0, k)$ be the smooth solution of the problem*

$$\begin{aligned} &(f'(B) - \xi)\frac{dB}{d\xi} = qk\left(\frac{\xi}{\eta_0}\right)^k, \qquad \eta_0 < \xi \le 0, \\ &B(\eta_0; \eta_0, k) = 0, \end{aligned} \tag{3.45}$$

where $\eta_0 = f'(0)$. Then $u = B(\xi; \eta_0, k)$ does not increase until it intersects with $\xi = f'(u)$ and $\lim_{k\to\infty} B(\xi; \eta_0, k) = \max\{a_1, R_1(\xi)\}$, $\xi \in (\eta_0, 0]$.

LEMMA 3.11. *Let $B(\xi; \eta_0, k)$ be the smooth solution of the problem*

$$\begin{cases} (f'(B) - \xi)\dfrac{dB}{d\xi} = qk\left(\dfrac{\xi}{\eta_0}\right)^k, \qquad \xi < 0, \\ B(0; \eta_0, k) = u^+. \end{cases} \tag{3.46}$$

Then for all $u^+ \in (u_2, a_2)$ or $u^+ \in (0, u_1)$, $f'(B(\xi; \eta_0, k) < \xi$, and therefore $B(\xi; \eta_0, k)$ does not decrease until it intersects with $\xi = f'(u)$. Furthermore, the limit of $B(\xi; \eta_0, k)$ as the reaction rate goes to infinity is

$$\lim_{k\to\infty} B(\xi; \eta_0, k) = u^+ \tag{3.47}$$

for all $\xi \in (f'(u^+), 0)$.

Note that $B(\eta_0; \eta_0, k)$ does not need to be zero, which is different from Lemma 3.10.

The proof of Lemmas 3.10 and 3.11 is similar to that of Lemmas 3.1 and 3.3. Using a method similar to that in Theorem 3.7, we can construct the solution of (2.2) and (2.4) for Case 3.2.1.

THEOREM 3.12. *When k is large enough, the Riemann problem of* (2.2) *and* (2.4) *for Case* 3.2.1 *has the unique admissible solution with the structure*

$$(u(\xi), z(\xi)) = \begin{cases} (\max\{u^-, R_1(\xi)\}, 1), & -\infty < \xi < \eta_0, \\ (B(\xi; \eta_0, k), (\frac{\xi}{\eta_0})^k), & \eta_0 < \xi < 0, \\ (C(\xi; \eta_0, k), 0), & 0 \le \xi < +\infty \end{cases} \tag{3.48}$$

and

$$\lim_{k\to\infty}(u(\xi), z(\xi)) = \begin{cases} (\max\{u^-, R_1(\xi)\}, 1), & -\infty < \xi \le \eta_0, \\ (C(\xi), 0), & \eta_0 < \xi < +\infty, \end{cases} \tag{3.49}$$

where $C(\xi)$ is the solution of

$$\begin{cases} (f'(C) - \xi)\frac{dC}{d\xi} = 0, \quad \eta_0 \le \xi < +\infty, \\ C(\eta_0) = a_1, \quad C(+\infty) = u^+. \end{cases} \tag{3.50}$$

Note here that $\eta_0 = \frac{f(a_1)-f(0)}{a_1-q} < f'(a_1)$.

Proof. Let $(u(\xi), z(\xi)) = (\max\{u^-, R_1(\xi)\}, 1)$ when $\xi \in (-\infty, \eta_0)$. Let $B_1(\xi;\eta_0,k)$ and $B_2(\xi;\eta_0,k)$ be the solutions of (3.45) and (3.46), respectively.

(i) If $\frac{f(u^+)-f(B_1(0;\eta_0,k))}{u^+-B_1(0;\eta_0,k)} \geq 0$, then we choose $u(\xi) = B_1(\xi;\eta_0,k)$ as the smooth solution of equation (3.45) when $\eta_0 < \xi < 0$. By Lemma 3.10, $\lim_{k\to\infty} B_1(\xi;\eta_0,k) = \max\{a_1, R_1(\xi)\}$, as $\xi \in (\eta_0, 0]$. Thus we have $u_1 < B(0;\eta_0,k) < u_2$, as k is large enough, due to $u_1 < a_1 < u_2$. Then the problem

$$\begin{cases} (f'(C)-\xi)\frac{dC}{d\xi} = 0, \quad 0 \leq \xi < +\infty, \\ C(0) = B_1(0;\eta_0,k), \quad C(+\infty;\eta_0,k) = u^+ \end{cases} \tag{3.51}$$

has a unique entropy solution since $\frac{f(u^+)-f(B_1(0;\eta_0,k))}{u^+-B_1(0;\eta_0,k)} \geq 0$.

(ii) If $\frac{f(u^+)-f(B_1(0;\eta_0,k))}{u^+-B_1(0;\eta_0,k)} < 0$, then we choose $u(\xi) = u^+$ as $\xi \geq 0$, $u(\xi) = B_1(\xi;\eta_0,k)$ as $\xi \in (\eta_0, \xi_k)$, and $u(\xi) = B_2(\xi;\eta_0,k)$ as $\xi \in (\xi_k, 0)$, where ξ_k is to be determined. Note that now $u^+ \notin (u_1, u_2)$ since $B_1(0;\eta_0,k) \in (u_1, u_2)$ and $\frac{f(B_1(0;\eta_0,k)-f(v)}{B_1(0;\eta_0,k)-v} > 0$ for all $v \in (u_1, u_2)$.

The summation of the integration of (3.45) from η_0 to ξ_k and the integration (3.46) from ξ_k to 0 results in

$$\begin{aligned} &[f(B_1(\xi_k;\eta_0,k)) - f(B_2(\xi_k;\eta_0,k))] \\ &- \xi_k[B_1(\xi_k;\eta_0,k) - B_2(\xi_k;\eta_0,k)] - [f(0) - f(u^+)] + q\eta_0\frac{k}{k+1} \\ &+ \int_{\eta_0}^{\xi_k} B_1(\xi;\eta_0,k)d\xi + \int_{\xi_k}^{0} B_2(\xi;\eta_0,k)d\xi = 0. \end{aligned} \tag{3.52}$$

Set

$$\begin{aligned} H(\xi) = &[f(B_1(\xi;\eta_0,k)) - f(B_2(\xi;\eta_0,k))] \\ &- \xi[B_1(\xi;\eta_0,k) - B_2(\xi;\eta_0,k)]. \end{aligned} \tag{3.53}$$

When $u^+ \in (0, u_1)$, we have

$$\frac{dH}{d\xi} = B_2(\xi;\eta_0,k) - B_1(\xi;\eta_0,k) < 0 \quad \text{for all } \xi \in (f'(u^+), 0) \tag{3.54}$$

and

$$H(0) = f(B_1(0;\eta_0,k)) - f(u^+) < 0. \tag{3.55}$$

In light of Lemmas 3.10 and 3.11, $\lim_{k\to\infty} B_1(\xi;\eta_0,k) = \max\{a_1, R_1(\xi)\}$ and $\lim_{k\to\infty} B_2(\xi;\eta_0,k) = u^+$ as $\xi \in (f'(u^+), 0)$. Since

$$\frac{f(a_1) - f(u^+)}{a_1 - u^+} > f'(u^+),$$

$H(\xi) > 0$ for sufficiently large k as $\xi > f'(u^+)$, and ξ is close to $f'(u^+)$. Therefore, there exists $\xi_k \in (f'(u^+), 0)$ such that $H(\xi_k) = 0$.

When $u^+ \in (u_2, a_2)$, we have

$$H(0) = f(B_1(0;\eta_0,k)) - f(u^+) > 0$$

and

$$H(\eta_0) = f(0) - f(B_2(\eta_0;\eta_0,k)) - \eta_0(0 - B_2(\eta_0;\eta_0,k) < 0;$$

there also exists $\xi_k \in (\eta_0, 0)$ such that $H(\xi_k) = 0$. Thus, define

$$\xi_k = \frac{f(B_1(\xi_k;\eta_0,k)) - f(B_2(\xi_k;\eta_0,k))}{B_1(\xi_k;\eta_0,k) - B_2(\xi_k;\eta_0,k)}. \tag{3.56}$$

By Lemmas 3.10 and 3.11, we have

$$\lim_{k\to\infty} B_1(\xi;\eta_0,k) = \max\{a_1, R_1(\xi)\}, \quad \lim_{k\to\infty} B_2(\xi;\eta_0,k) = u^+.$$

Then, as k is sufficiently large, $B_1(\xi_k;\eta_0,k)$ is close to $\max\{a_1, R_1(\xi_k)\}$ and $B_2(\xi_k;\eta_0,k)$ is close to u^+. Therefore, the discontinuity of u at $\xi = \xi_k$ satisfies the Oleinik entropy condition (2.7). By (3.52), we conclude that this solution satisfies (2.2) and (2.4) in the sense of distributions. Thus we have constructed the admissible solution. The uniqueness is obvious.

Using Lemmas 3.10 and 3.11, the limit of the solution is obtained. □

Case 3.2.2. $u^+ \in [a_2, +\infty)$. Set

$$(u(\xi), z(\xi)) = \begin{cases} (\max\{u^-, R_1(\xi)\}, 1), & \xi \in (-\infty, \eta_k), \\ (B(\xi;\eta_k,k), (\frac{\xi}{\eta_k})^k), & \xi \in (\eta_k, 0], \\ (u^+, 0), & \xi \in (0, +\infty), \end{cases} \tag{3.57}$$

where η_k is to be determined, and $B(\xi;\eta_k,k)$ satisfies

$$\begin{aligned} &(f'(B) - \xi)\frac{dB}{d\xi} = qk\left(\frac{\xi}{\eta_k}\right)^k, \quad \eta_k < \xi < 0, \\ &B(0;\eta_k,k) = u^+. \end{aligned} \tag{3.58}$$

Since $u^+ \geq a_2$, the Riemann problem of (2.2) and (2.4) does not have the same kind of solution as that in Case 3.2.1. This implies that $\eta_k < \eta_0$. Denote $\eta^- = f'(u^-)$. The summation of the integration of (3.3) from η^- to η_k when $q = 0$ and the integration from η_k to 0 when $q > 0$ is

$$\begin{aligned} &f(\max\{u^-, R_1(\eta_k)\}) - f(B(\eta_k;\eta_k,k)) - \eta_k(\max\{u^-, R_1(\eta_k)\} - B(\eta_k;\eta_k,k)) \\ &+ f(u^+) - f(u^-) + \eta^- u^- + \int_{\eta^-}^{\eta_k} \max\{u^-, R_1(\xi)\}d\xi + \int_{\eta_k}^{0} B(\xi;\eta_k,k)d\xi + \frac{kq\eta_k}{k+1} = 0. \end{aligned} \tag{3.59}$$

Set

$$\begin{aligned} H(\eta) = {} & f(B(\eta;\eta,k)) - f(\max\{u^-, R_1(\eta)\}) \\ & - \eta[B(\eta;\eta,k) - \max\{u^-, R_1(\eta)\}]. \end{aligned} \tag{3.60}$$

Then we claim that there exists $\eta_k < \eta_0$ such that $H(\eta_k) = 0$.

Draw a line $w - f(u^-) = f'(u^-)(u - (u^- + q))$. It has an intersection point, denoted by $(b_1, f(b_1))$, $b_1 > a_2$, with $w = f(u)$. Then we prove the above claim using two cases.

(i) $u^+ \in [a_2, b_1)$. Then we can find $u^* \in (u^-, 0]$ such that

$$f'(u^*) = \frac{f(u^+) - f(u^*)}{u^+ - (u^* + q)} =: \eta^*. \tag{3.61}$$

In light of Lemma 3.11, for sufficiently large k, the solution $B(\eta; \eta, k)$ of (3.58) is close to u^+. Therefore,

$$H(\eta^*) = f(u^*) - f(B(\eta^*; \eta^*, k)) - \eta^*(u^* - B(\eta^*; \eta^*, k)) > 0. \tag{3.62}$$

On the other hand,

$$H(\eta^-) = f(u^-) - f(B(\eta^-; \eta^-, k)) - \eta^-(u^- - B(\eta^-; \eta^-, k)) < 0. \tag{3.63}$$

Hence, there exists $\eta_k \in (\eta^-, \eta^*)$ such that $H(\eta_k) = 0$.

(ii) $u^+ \in [b_1, +\infty)$. Now we set

$$\eta^* = \frac{f(u^+) - f(u^-)}{u^+ - (u^- + q)}. \tag{3.64}$$

Then we assert

$$\eta^- > \eta^* > \eta^+ := f'(u^+). \tag{3.65}$$

Making use of (3.59), we can conclude that

$$H(\eta^+) < 0 \quad \text{and} \quad H(\eta^-) > 0. \tag{3.66}$$

Therefore, there also exists $\eta_k \in (\eta^+, \eta^-)$ such that $H(\eta_k) = 0$.

Thus we construct an admissible solution of (2.2) and (2.4) for Case 3.2.2. The uniqueness is obvious. Therefore, we summarize to obtain the following theorem.

THEOREM 3.13. *The Riemann problem for Case* 3.2.2 *has a unique admissible solution with the structure*

$$(u(\xi), z(\xi)) = \begin{cases} (\max\{u^-, R_1(\xi)\}, 1), & \xi \in (-\infty, \eta_k), \\ (B(\xi; \eta_k, k), (\frac{\xi}{\eta_k})^k), & \xi \in (\eta_k, 0), \\ (u^+, 0), & \xi \in [0, +\infty) \end{cases} \tag{3.67}$$

and

$$\lim_{k \to +\infty} (u(\xi), z(\xi)) = \begin{cases} (\max\{u^-, R_1(\xi)\}, 1), & \xi \in (-\infty, \eta^*), \\ (u^+, 0), & \xi \in (\eta^*, +\infty). \end{cases} \tag{3.68}$$

4. Entropy condition for combustion waves of the CJ model. In this section, we will propose the entropy condition for combustion waves of the CJ model (1.2) with nonconvex fluxes $f(u)$ by taking into account the limit behavior of the solutions $(u(\xi), z(\xi))$ of (2.2) and (2.4) as the reaction rate goes to infinity. As we discussed in section 3, the Riemann solutions are essentially classified into two kinds: noncombustion solutions and combustion solutions. The combustion solutions have two types.

(i) For Type 1, as shown in Cases 3.1.2 and 3.2.2, the gas is ignited through a shock at $\xi = \eta_k$. The position of the reaction wave front depends on the reaction rate

k. The solution u is increasing as $\xi < \eta_k$, while it is decreasing as $\xi > \eta_k$. There is a von Neumann spike $u = B(\eta_k; \eta_k, k)$ on the curve of $u = u(\xi)$ in a neighborhood of $\xi = \eta_k$. Denote $\bar{\eta} = \lim_{k\to+\infty} \eta_k$, $\bar{u}(\xi) = \lim_{k\to+\infty} B(\xi; \eta_k, k)$ for $\xi \in (\bar{\eta}, 0)$ and $\bar{u}_R = \lim_{k\to+\infty} B(\eta_k; \eta_k, k)$.

Then, at $\xi = \bar{\eta}$, we have the relation

$$\bar{\eta} = \frac{f(\bar{u}_r) - f(\bar{u}_l)}{\bar{u}_r - \bar{u}_l - q} < 0, \qquad f'(\bar{u}_l) \geq \bar{\eta} \geq f'(\bar{u}_r), \tag{4.1}$$

and $\bar{u}_R > \bar{u}_r$ such that

$$\frac{f(\bar{u}_R) - f(\bar{u}_l)}{\bar{u}_R - \bar{u}_l} = \frac{f(\bar{u}_r) - f(\bar{u}_l)}{\bar{u}_r - \bar{u}_l - q} \leq \frac{f(u) - f(\bar{u}_l)}{u - \bar{u}_l} \quad \text{for all } u \in (\bar{u}_l, \bar{u}_R), \tag{4.2}$$

where $\bar{u}_r = \bar{u}(\bar{\eta} + 0)$, $\bar{u}_l = \bar{u}(\bar{\eta} - 0)$.

(ii) For Type 2, as shown in Cases 3.1.1 and 3.2.1, the gas will burn when its temperature reaches the ignition point continuously at $\xi = \eta_0$ or jumps over the ignition point through a shock at $\xi = \eta^*$. The position of the reaction wave front $\xi = \eta_0$ (or $\xi = \eta^*$) does not depend on the reaction rate k. In the neighborhood of $\xi = \eta_0$ or $\xi = \eta^*$, u is increasing. There is no von Neumann spike on the curve of $u = u(\xi)$ in the neighborhood of $\xi = \eta_0$ or $\xi = \eta^*$. Denote $\bar{\eta} = \eta_0$ or η^*, $\bar{u}(\xi) = \lim_{k\to+\infty} B(\xi; \bar{\eta}, k)$ for $\xi \in (\bar{\eta}, 0)$, and $\bar{u}_R = \lim_{k\to+\infty} B(\bar{\eta}; \bar{\eta}, k)$. Then, at $\xi = \bar{\eta}$, we have the relation

$$\bar{\eta} = \frac{f(\bar{u}_r) - f(\bar{u}_l)}{\bar{u}_r - \bar{u}_l - q} < 0, \qquad \bar{\eta} \leq f'(\bar{u}_r), \tag{4.3}$$

and $\bar{u}_R \in [0, \bar{u}_r)$ such that

$$\frac{f(\bar{u}_R) - f(\bar{u}_l)}{\bar{u}_R - \bar{u}_l} = \frac{f(\bar{u}_r) - f(\bar{u}_l)}{\bar{u}_r - \bar{u}_l - q} \geq \frac{f(u) - f(\bar{u}_l)}{u - \bar{u}_l} \quad \text{for all } u \in (\bar{u}_l, \bar{u}_r), \tag{4.4}$$

where $\bar{u}_r = \bar{u}(\bar{\eta} + 0)$, $\bar{u}_l = \bar{u}(\bar{\eta} - 0)$.

DEFINITION 4.1. *For the CJ combustion model* (1.2), *we define the limit of the interface between the unburned and reaction states in Type* I *as a generalized detonation wave and in Type* II *as the generalized deflagration wave.*

From this definition, we extract the following entropy condition on combustion waves for the nonconvex CJ combustion model (1.2).

ENTROPY CONDITION. *Let $x = x(t)$ be a combustion wave of the CJ combustion model* (1.2). *Let $u_l = u(x(t) - 0, t)$ and $u_r = u(x(t) + 0, t)$ be the limit values of u in the wave front and the wave back, respectively. Then*

(1) *$x = x(t)$ is a generalized CJ deflagration wave if*

$$\frac{dx}{dt} = \frac{f(u_r) - f(u_l)}{u_r - u_l - q} < 0, \quad \frac{dx}{dt} \leq f'(u_r), \tag{4.5}$$

and there exists $u_R \in [0, u_r)$ such that

$$\frac{f(u_R) - f(u_l)}{u_R - u_l} = \frac{f(u_r) - f(u_l)}{u_r - u_l - q} \geq \frac{f(u) - f(u_l)}{u - u_l} \quad \textit{for all } u \in (u_l, u_r); \tag{4.6}$$

(2) *$x = x(t)$ is a generalized detonation wave if*

$$\frac{dx}{dt} = \frac{f(u_r) - f(u_l)}{u_r - u_l - q} < 0, \qquad f'(u_l) \geq \frac{dx}{dt} \geq f'(u_r), \tag{4.7}$$

and there exists $u_R \in (u_r, +\infty)$ *such that*

$$\frac{f(u_R)-f(u_l)}{u_R-u_l}=\frac{f(u_r)-f(u_l)}{u_r-u_l-q}\le\frac{f(u)-f(u_l)}{u-u_l}\quad \textit{for all } u\in(u_l,u_R). \tag{4.8}$$

Furthermore the detonation wave is a CJ detonation wave if $\frac{dx}{dt}=f'(u_r)$*; otherwise, it is a strong detonation wave.*

As we have seen, this entropy condition for the nonconvex CJ combustion model (1.2) inherits the essential difference between detonation and deflagration waves in that the former contains a von Neumann spike in the finite reaction rate region but the latter does not, which reflects the intrinsic feature of combustion waves in gas dynamics (cf. [1, 22]). With this, we can improve the results in [18] greatly to justify the (entropy) solutions of the Riemann problem for (1.2). Actually this entropy condition can be used in the construction of two-dimensional Riemann problems for the counterpart of (1.2), the generalization of (1.2) in two-dimensions; see [20]. Therefore, to some extent, this entropy condition lays a foundation and gives insight into the study of the structure of multidimensional combustion solutions.

Acknowledgments. The authors are grateful to Professor Tong Zhang for stimulating discussions, and also thank the referees for pointing out unclear statements and errors and improving the presentation of this paper. The second author is also grateful to the Institute of Mathematics, the Hebrew University of Jerusalem, whose hospitality and Professor Matania Ben-Artzi's invitation are greatly appreciated.

REFERENCES

[1] R. Courant and K. O. Friedrichs, *Supersonic flow and shock waves*, Interscience, New York, 1948.
[2] W. Fickett, *Detonation in miniature*, Amer. J. Phys., 47 (1979), pp. 1050–1059.
[3] C.-H. Hsu and S.-S. Lin, *Some qualitative properties of the Riemann problem in gas dynamical combustion*, J. Differential Equations, 140 (1997), pp. 10–43.
[4] P. D. Lax, *The multiplicity of eigenvalues*, Bull. Amer. Math. Soc. (N.S.), 6 (1982), pp. 213–214.
[5] A. Levy, *On Majda's model for dynamic combustion*, Comm. Partial Differential Equations, 17 (1992), pp. 657–698.
[6] J. Li, *On the uniqueness and existence problem for a multidimensional reacting and convection system*, J. London Math. Soc., 62 (2000), pp. 473–488.
[7] T. Li, *Time-asymptotic limit of solutions of a combustion problem*, J. Dynam. Differential Equations, 10 (1998), pp. 577–604.
[8] T. Li, *Rigorous asymptotic stability of a Chapman-Jouguet detonation wave in the limit of small resolved heat release*, Combust. Theory Model., 1 (1997), pp. 259–270.
[9] T. Li, *On the initiation problem for a combustion model*, J. Differential Equations, 112 (1994), pp. 351–373.
[10] T. Li, *On the Riemann problem for a combustion model*, SIAM J. Math. Anal., 24 (1993), pp. 59–75.
[11] T.-P. Liu and L. A. Ying, *Nonlinear stability of strong detonations for a viscous combustion model*, SIAM J. Math. Anal., 26 (1995), pp. 519–528.
[12] T.-P. Liu and S.-H. Yu, *Nonlinear stability of weak detonation waves for a combustion model*, Comm. Math. Phys., 204 (1999), pp. 551–586.
[13] T.-P. Liu and T. Zhang, *A scalar combustion model*, Arch. Ration. Mech. Anal., 114 (1991), pp. 297–312.
[14] A. Majda, *A qualitative model for dynamic combustion*, SIAM J. Appl. Math., 41 (1981), pp. 70–93.
[15] R. R. Rosales and A. Majda, *Weakly nonlinear detonation waves*, SIAM J. Appl. Math., 43 (1983), pp. 1086–1118.
[16] D. Tan and T. Zhang, *Riemann problem for the selfsimilar ZND-model in gas dynamical combustion*, J. Differential Equations, 95 (1992), pp. 331–369.

[17] L. Ying and Z. Teng, *Riemann problem for a reaction and convection hyperbolic system*, Approx. Theory Appl., 1 (1984), pp. 95–122.
[18] P. Zhang and T. Zhang, *The Riemann problem for scalar CJ-combustion model without convexity*, Discrete Contin. Dynam. Systems, 1 (1995), pp. 195–206.
[19] P. Zhang and T. Zhang, *The Riemann problem for nonconvex combustion model from ZND to CJ theory*, in Advances in Nonlinear Partial Differential Equations and Related Areas (Beijing, 1997), World Scientific, River Edge, NJ, 1998, pp. 379–398.
[20] P. Zhang and T. Zhang, *The Two-Dimensional Riemann Problem for a Simplified Combustion Model*, in preparation, 2001.
[21] T. Zhang, *The Riemann problem for combustion*, Contemp. Math., 100 (1989), pp. 111–124.
[22] T. Zhang and Y. Zheng, *Riemann problem for gasdynamic combustion*, J. Differential Equations, 77 (1989), pp. 203–230.

SIAM J. MATH. ANAL.
Vol. 34, No. 3, pp. 700–718

WIGNER MEASURE AND THE SEMICLASSICAL LIMIT OF SCHRÖDINGER–POISSON EQUATIONS*

PING ZHANG†

Abstract. The wave function $\{\psi^\epsilon(t,x)\}$ of single particle approximation, which is used in the study of quantum transportation in some semiconductive devices, satisfies Schrödinger–Poisson equations. It is well known that the Wigner transformation $f^\epsilon(t,x,\xi)$ of the corresponding wave function $\psi^\epsilon(t,x)$ satisfies the so-called Wigner–Poisson equations. We prove here that in any space dimension, with the initial data of the form $\sqrt{\rho_0^\epsilon(x)}\exp(\frac{i}{\epsilon}S^\epsilon(x))$ to the wave function, and before the formation of vortices, the Wigner measure $f(t,x,\xi)$, which is the weak limit of $f^\epsilon(t,x,\xi)$ as the normalized Planck constant ϵ approaches 0, satisfies Vlasov–Poisson equations, and the limits of the quantum density and momentum to the Schrödinger–Poisson equations satisfy the pressureless Euler–Poisson equations.

Key words. Schrödinger–Poisson, Euler–Poisson, Vlasov–Poisson, Wigner measure, Wigner transformation, pseudodifferential operator

AMS subject classifications. 35L65, 81

PII. S0036141001393407

1. Introduction.

In this paper, we consider the local-in-time semiclassical limit of Schrödinger–Poisson equations in any space dimension. The wave function $\{\psi^\epsilon(t,x)\}$ of single particle approximation, which is used in the study of quantum transportation in some semiconductive devices, satisfies Schrödinger–Poisson equations,

$$\begin{cases} i\epsilon\partial_t\psi^\epsilon = -\frac{\epsilon^2}{2}\triangle\psi^\epsilon + V^\epsilon\psi^\epsilon, \qquad x\in\mathbb{R}^d, \quad t\geq 0, \\ \psi^\epsilon(t=0,x) = \sqrt{\rho_0^\epsilon(x)}\exp(\frac{i}{\epsilon}S^\epsilon(x)), \end{cases} \tag{1.1}$$

where ϵ is the normalized Planck constant, and the potential $V^\epsilon(t,x)$ is assumed to be given self-consistently by Poisson's equation,

$$-\triangle V^\epsilon = \rho^\epsilon - b(x), \quad \rho^\epsilon(t,x) = |\psi^\epsilon(t,x)|^2, \tag{1.2}$$

where V^ϵ and $\nabla_x V$ vanish as $|x|\to\infty$, and the function $b(x)$ denotes the doping profile in the semiconductor applications, and, in general, it denotes a fixed background charge. One may see [17] or [5] for more physical explanations..

As was commented in [5], it is a fundamental principle in quantum mechanics that when the time and distance scales are large enough relative to the Planck's constant, the quantum density, $|\psi^\epsilon|^2$, and the quantum momentum, $\epsilon\mathrm{Im}(\overline{\psi^\epsilon}\nabla\psi^\epsilon)$, will approximately obey the laws of classical, Newtonian mechanics. And the quantum-mechanical pressure disappears in the semiclassical limit, and Euler equations for an isentropic compressible flow are formally recovered from the nonlinear Schrödinger equation. In one space dimension with $V^\epsilon = |\psi^\epsilon|^2 - 1$, the global character of the semiclassical limit was established by Jin, Levermore, and McLaughlin [10], [11] using

*Received by the editors August 7, 2001; accepted for publication (in revised form) June 20, 2002; published electronically January 7, 2003. This work was partially supported by the Chinese Youth Foundation, the innovation grant from the Chinese Academy of Sciences, and the Austrian START Project.

http://www.siam.org/journals/sima/34-3/39340.html

†Academy of Mathematics and System Sciences, The Chinese Academy of Sciences, Beijing 100080, China (zp@math03.math.ac.cn).

the inverse scattering method; for $d \geq 2$, $V^\epsilon = f(|\psi^\epsilon|^2)$ with $f' > 0$, before the formation of singularities in the limit system, Grenier [9] solved the limit by applying the symmetric hyperbolic equation theory. The semiclassical limit of the general modified nonlinear Schrödinger equation (see [3]) can be compared to the similar strategy of Grenier's. But we will see that this method does not work for the semiclassical limit of the Schrödinger–Poisson equations if we follow the idea of [9], as the resulting limit equations are not a symmetric hyperbolic system; thus the standard energy estimate which works well there fails here.

Traditional method suggests that, at least for short time, the wave function $\psi^\epsilon(t,x)$ for (1.1) will be of the following form:

$$\psi^\epsilon(t,x) = \sqrt{\rho^\epsilon(t,x)} \exp\left(\frac{i}{\epsilon} S^\epsilon(t,x)\right). \tag{1.3}$$

Then by substituting (1.3) to (1.1) and separating the real and imaginary part in (1.1), the irrotational flow equations

$$\begin{cases} \partial_t \rho^\epsilon + \operatorname{div} J^\epsilon = 0, \\ \partial_t J^\epsilon + \operatorname{div}\left(\frac{J^\epsilon \otimes J^\epsilon}{\rho^\epsilon}\right) + \rho^\epsilon \nabla V^\epsilon = \frac{\epsilon^2}{2} \rho^\epsilon \nabla\left[\frac{1}{\sqrt{\rho^\epsilon}} \triangle \sqrt{\rho^\epsilon}\right], \end{cases} \tag{1.4}$$

where $J^\epsilon = \rho^\epsilon \nabla S^\epsilon = \epsilon \operatorname{Im}(\overline{\psi^\epsilon} \nabla \psi^\epsilon)$ and $\otimes$ denotes the tensor product of vectors (see the notation at the end of the introduction), are obtained in [12]. Equations (1.4) represent a fluid dynamic formulation of (1.1) and are known as Madelung's fluid equations [15]. The semiclassical limit of (1.1) just means the vanishing dispersion limit to (1.4). Unfortunately, we can do almost nothing for the vanishing dispersion limit due to the strong singularity on the set $\{(t,x) | \rho^\epsilon(t,x) = 0\}$.

Motivated by recent work of Brenier [1], where the author proved the local-in-time convergence of the scaled Vlasov–Poisson equations to the incompressible Euler equations, we are going to use the Wigner measure approach to study the semiclassical limit of (1.1). In 1932, Wigner [23] introduced the following transformation in quantum mechanics:

$$f^\epsilon(t,x,\xi) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\xi y} \psi^\epsilon\left(t, x + \frac{\epsilon y}{2}\right) \overline{\psi^\epsilon\left(t, x - \frac{\epsilon y}{2}\right)}\, dy. \tag{1.5}$$

Then trivial calculation shows that $f^\epsilon(t,x,\xi)$ satisfies the equation

$$\begin{cases} \partial_t f^\epsilon + \xi \nabla f^\epsilon + \theta[V^\epsilon] f^\epsilon = 0, \\ f^\epsilon(t=0,x,\xi) = f_I^\epsilon(x,\xi), \end{cases} \tag{1.6}$$

where $\theta[V^\epsilon] f^\epsilon(t,x,\xi)$ is the pseudodifferential operator

$$\theta[V^\epsilon] f^\epsilon(t,x,\xi) = \frac{i}{(2\pi)^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{V^\epsilon(t, x + \frac{\epsilon y}{2}) - V^\epsilon(t, x - \frac{\epsilon y}{2})}{\epsilon} f^\epsilon(t,x,\eta) e^{-i(\xi-\eta)y}\, d\eta\, dy. \tag{1.7}$$

Formally passing $\epsilon \to 0$ in (1.6), we get

$$\begin{cases} \partial_t f + \xi \nabla_x f - E \nabla_\xi f = 0, \qquad E = \nabla \triangle^{-1}(b(x) - \rho), \\ f(t=0,x,\xi) = f_0(x,\xi), \end{cases} \tag{1.8}$$

where $f(t,x,\xi)$ is the weak limit of $f^\epsilon(t,x,\xi)$, which is a nonnegative Radon measure (see [14] or [7] for more details). Recently we [25] rigorously justified the limit

from (1.6) to (1.8) in one space dimension with very general initial data to the wave function.

On the other hand, with the initial data of the form $\psi_0^\epsilon(x) = \sqrt{\rho_0^\epsilon(x)}\exp(\frac{i}{\epsilon}S^\epsilon(x))$, and assuming that $\{\rho_0^\epsilon(x)\}, \{\nabla S^\epsilon(x)\}$ converge to $\rho_0(x)$ and $u_0(x)$ in some sense (see assumption (A3) below), we can easily calculate that the corresponding Wigner measure $f_0(x,\xi)$ is $\rho_0(x)\delta(\xi - u_0(x))$. Taking this $f_0(x,\xi)$ as initial data for (1.8), formally we can expect that there is a positive constant T^* such that (1.8) has a solution of the form $\rho(t,x)\delta(\xi - u(t,x))$ for $t < T^*$, and $(\rho(t,x), u(t,x))$ is smooth for $t < T^*$. Then by the velocity average to (1.8), we find

$$\left\{\begin{array}{l} \partial_t\rho + \mathrm{div}(\rho u) = 0, \\ \partial_t(\rho u) + \mathrm{div}(\rho u\otimes u) + \rho\nabla V = 0, \qquad \triangle V = b(x) - \rho. \end{array}\right. \tag{1.9}$$

In the following, we will give a rigorous justification to the above formal explanation. It should be noticed that this idea has been used by Gasser and Markowich [4] in the study of the semiclassical limit of the linear Schrödinger equation, where the limit from (1.6) to (1.8) was proved in [14].

Before the presentation of the main result of this paper, let us first make the following assumptions:

(A1) $b(x), \sqrt{\rho_0^\epsilon(x)}, S^\epsilon(x) \in H^s(\mathbb{R}^d)$ for $s \in \mathbb{Z}^+$ and $s > \frac{d}{2} + 2$;

(A2) $b(x) \in L^1(\mathbb{R}^d)$, $\rho_0^\epsilon(x)$ is uniformly bounded in $L^1(\mathbb{R}^d)\cap L^2(\mathbb{R}^d)$, and $\nabla\sqrt{\rho_0^\epsilon}(x)$ and $\sqrt{\rho_0^\epsilon}\nabla S^\epsilon(x)$ are uniformly bounded in $L^2(\mathbb{R}^d)$;

(A3) $\nabla\triangle^{-1}(\rho_0^\epsilon - \rho_0)(x) \to 0$, $\sqrt{\rho_0^\epsilon}(\nabla S^\epsilon - u_0)(x) \to 0$ in $L^2(\mathbb{R}^d)$, and $\rho_0(x) \in H^{s-1}(\mathbb{R}^d), u_0(x) \in H^s(\mathbb{R}^d)$, with $s \in \mathbb{Z}^+$ and $s > \frac{d}{2} + 2$;

(A4) when $d = 1$, 2, we further need that

$$\int_{\mathbb{R}^d} \rho_0^\epsilon(x)\,dx = \int_{\mathbb{R}^d} b(x)\,dx = \int_{\mathbb{R}^d} \rho_0(x)\,dx,$$
$$\int_{\mathbb{R}^d} |x||\rho_0^\epsilon(x) - b(x)|\,dx < C,$$

with C an ϵ-independent positive constant.

From now on, let us denote by (ρ, u) the local smooth solution to the following equations:

$$\left\{\begin{array}{l} \partial_t\rho + \mathrm{div}(\rho u) = 0, \\ \partial_t u + u\nabla u + \nabla V = 0, \quad \triangle V = b(x) - \rho, \\ \rho(t=0,x) = \rho_0, \quad u(t=0,x) = u_0, \end{array}\right. \tag{1.10}$$

where V and $\nabla_x V$ vanish as $|x| \to \infty$. Notice that any smooth solution of (1.10) must be a smooth solution of (1.9). Because of the degeneracy of the second equation of (1.9) on the set $\{(t,x) : \rho(t,x) = 0\}$, we will consider (1.10) as the limit system instead of (1.9).

For the convenience of the reader, let us recall the test function space from [14],

$$\mathcal{A} = \left\{\phi \in C_c^\infty(\mathbb{R}^d_x \times \mathbb{R}^d_\xi)|, \quad (\mathcal{F}_\xi\phi)(x,\eta) \in L^1(\mathbb{R}^d_\eta, C_c(\mathbb{R}^d_x))\right\}, \tag{1.11}$$

with the norm

$$\|\phi(x,\xi)\|_{\mathcal{A}} = \int_{\mathbb{R}^d} \sup_x |(\mathcal{F}_\xi\phi)(x,\eta)|\,d\eta, \tag{1.12}$$

where $(\mathcal{F}_\xi \phi)(x,\eta)$ is the Fourier transformation of $\phi(x,\xi)$ with respect to ξ.

DEFINITION 1.1. *We call* $f^\epsilon(t,x,\xi)$ *the* Wigner transformation *of* $\psi^\epsilon(t,x)$ *for any fixed* t. *The corresponding weak limit* $f(t,x,\xi)$, *which we will detail in what follows, of* $f^\epsilon(t,\cdot,\cdot)$ *as* $\epsilon \to 0$ *is called the* Wigner measure *of* $\psi^\epsilon(t,x)$.

Remark 1.1. The Wigner measure $f(t,x,\xi)$ is in fact a nonnegative Radon measure for any fixed time t. See [14] and [7] for more details. Moreover, $f(t,x,\xi)$ has a very close relation with the H-measure of $\{\psi^\epsilon(t,x))\}$ (see [20], [6], or [14]).

Then we have the following theorem.

THEOREM 1.1. *Let* $(\rho_0^\epsilon(x), S^\epsilon(x)), (\rho_0(x), u_0(x))$ *satisfy* (A1)–(A4), *and let* $\psi^\epsilon(t,x)$, $(\rho(t,x), u(t,x))$ *be the solutions of* (1.1)–(1.2) *and* (1.10), *respectively. Then there exists a positive constant* T^* *such that for all* $T < T^*$, $\rho(t,x) \in L^\infty([0,T], H^{s-1}(\mathbb{R}^d))$, $u(t,x) \in L^\infty([0,T], H^s(\mathbb{R}^d))$. *Moreover, for any fixed* $t < T^*$, *the following hold:*

(1)

$$f^\epsilon(t,x,\xi) \rightharpoonup f(t,x,\xi) =: \rho(t,x)\delta(\xi - u(t,x)) \quad in \quad \mathcal{A}'(\mathbb{R}^{2d}), \tag{1.13}$$

$$|\psi^\epsilon(t,x)|^2 \rightharpoonup \rho(t,x) \quad in \quad \mathcal{M}^+(\mathbb{R}^d), \tag{1.14}$$

$$\epsilon \mathrm{Im}\left(\overline{\psi^\epsilon}(t,x)\nabla\psi^\epsilon(t,x)\right) \rightharpoonup (\rho u)(t,x) \quad in \quad \mathcal{M}(\mathbb{R}^d), \tag{1.15}$$

as ϵ *tends to* 0, *and* T^* *is the first time such that*

$$\lim_{T \to T^*} \|\nabla u(t,\cdot)\|_{L^\infty([0,T]\times\mathbb{R}^d)} = \infty. \tag{1.16}$$

(2) $f(t,x,\xi) \in Lip([0,T^*), H^{-s}(\mathbb{R}^{2d}))$ *for* $s > d+1$ *and* $f(t,x,\xi)$ *satisfies* (1.8) *on* $[0,T^*) \times \mathbb{R}^{2d}$ *in the sense of distribution.*

Remark 1.2. When $d = 1$ and $\rho_0^\epsilon(x)$, $S^\epsilon(x)$ are smooth, we can prove that the Wigner measure $f(t,x,\xi)$ globally solves the one-dimensional Vlasov–Poisson equations (see [25] for more details). Then by (3.35) and the proof of (3.41), up to a subsequence of $\{\psi^\epsilon\}$, which we denote by $\{\psi^{\epsilon_j}\}$ for any fixed $t \geq T^*$, there holds

$$\begin{aligned} &f^\epsilon(t,x,\xi) \rightharpoonup f(t,x,\xi) \quad \text{in} \quad \mathcal{A}'(\mathbb{R}^{2d}), \\ &\epsilon \mathrm{Im}\left(\overline{\psi^\epsilon}(t,x)\nabla\psi^\epsilon(t,x)\right) \rightharpoonup \int_{\mathbb{R}} \xi f(t,x,d\xi) \quad \text{in} \quad \mathcal{M}(\mathbb{R}^d) \end{aligned}$$

as $\epsilon \to 0$, where $f(t,x,\xi)$ is the global weak solution of the one-dimensional Vlasov–Poisson equations

$$\begin{cases} \partial_t f + \xi \partial_\xi - E\partial_\xi f = 0, \quad x \in \mathbb{R}, t \geq 0, \\ \partial_x E = b(x) - \int_{\mathbb{R}} f(t,x,d\xi), \\ f(t=0,x,\xi) = \rho_0(x)\delta(\xi - u_0(x)). \end{cases} \tag{1.17}$$

But from [11], we cannot expect that the limits of the quantum density and momentum to the Schrödinger–Poisson equations satisfy (1.9) after the formation of singularities in the limit system.

Remark 1.3. Recently Lin and Zhang [13] rigorously proved that before the formation of singularities in the limit system, the limit equation obtained by the hydrodynamic limit of Ginzburg–Landau wave vortices is again (1.10).

Notation used throughout this paper. Let $a = (a_1, a_2, \ldots, a_d)$, $b = (b_1, b_2, \ldots, b_d)$ be two vectors; then we denote

$$a \otimes b = (a_i b_j)_{n\times n}.$$

Let $M^k(x) = \big(m^k_{ij}(x)\big)_{n\times n}$, $k = 1, 2$, be two $C^1(\mathbb{R}^d)$ matrix functions; then

$$M^1 : M^2 = \sum_{i,j=1}^{d} m^1_{ij} m^2_{ij}, \quad \nabla : M^1(x) = \begin{pmatrix} \mathrm{div}\vec{m}^1_1(x) \\ \cdot \\ \cdot \\ \cdot \\ \mathrm{div}\vec{m}^1_d(x) \end{pmatrix},$$

with $\vec{m}^1_i(x) = \big(m^1_{i1}(x), m^1_{i2}(x), \ldots, m^1_{id}(x)\big)$.

We denote by $\mathcal{M}\ (\mathbb{R}^d)$ the Radon measure space over $\mathbb{R}^d$, by $\hat{f}(\xi)$ the Fourier transform of $f(x)$, by $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ the multi-index, $D_x = \frac{1}{i}\nabla_x$, and by the letter C a uniform constant, which may change from line to line.

2. Preliminaries. In this section, we will present some uniform estimates to the smooth solution of the quantum Schrödinger–Poisson equations and prove the local existence of the smooth solution to (1.10).

LEMMA 2.1. *Let $(\rho^\epsilon_0(x), S^\epsilon(x))$ satisfy* (A1); *then for any fixed ϵ,* (1.1)–(1.2) *has a unique global smooth solution $\psi^\epsilon(t,x) \in L^\infty([0,T], H^s(\mathbb{R}^d))$ for any $T < \infty$. Moreover, the following hold:*

(1)

$$\|\psi^\epsilon(t,\cdot)\|_{L^2} = \|\rho^\epsilon_0\|^{\frac{1}{2}}_{L^1}; \tag{2.1}$$

(2)

$$\frac{d}{dt}\left\{\frac{\epsilon^2}{2}\|\nabla\psi^\epsilon(t,\cdot)\|^2_{L^2} + \|\nabla V^\epsilon(t,\cdot)\|^2_{L^2}\right\} = 0. \tag{2.2}$$

Proof. First, for any fixed ϵ, by modifying the method in [8] or [2], we easily obtain that (1.1) has a global smooth solution $\psi^\epsilon(t,x) \in L^\infty([0,T], H^s(\mathbb{R}^d))$ for any $T < \infty$, with initial data $\sqrt{\rho^\epsilon_0(x)}\exp(\frac{i}{\epsilon}S^\epsilon(x))$. Details are omitted.

By multiplying $\overline{\psi^\epsilon}(t,x)$ by both sides of (1.1), then integrating with respect to x over $\mathbb{R}^d$ and taking integration by parts, we obtain part (1) of Lemma 2.1.

Next, let us multiply $\partial_t\overline{\psi^\epsilon}(t,x)$ by both sides of (1.1), then integrating with respect to x and integrating by parts again, we have

$$i\epsilon\int_{\mathbb{R}^d}|\partial_t\psi^\epsilon|^2\,dx = \frac{\epsilon^2}{2}\int_{\mathbb{R}^d}\nabla\psi^\epsilon\partial_t\nabla\overline{\psi^\epsilon}\,dx + \int_{\mathbb{R}^d}V^\epsilon\psi^\epsilon\partial_t\overline{\psi^\epsilon}\,dx. \tag{2.3}$$

Taking the real part of (2.3), we find

$$\begin{aligned}\frac{\epsilon^2}{2}\frac{d}{dt}\int_{\mathbb{R}^d}|\nabla\psi^\epsilon(t,\cdot)|^2\,dx &= -\int_{\mathbb{R}^d}V^\epsilon\partial_t|\psi^\epsilon(t,\cdot)|^2\,dx \\ &= \int_{\mathbb{R}^d}V^\epsilon\partial_t\triangle V^\epsilon\,dx = -\frac{d}{dt}\int_{\mathbb{R}^d}|\nabla V^\epsilon(t,\cdot)|^2\,dx,\end{aligned} \tag{2.4}$$

where we used (1.2) in the next-to-last step of (2.4). Then (2.2) is a direct consequence of (2.4), which completes the proof of the lemma. ∎

Remark 2.1. By (A2) and (A4), we can prove that $\{\nabla V^\epsilon(0,x)\}$ is uniformly bounded in $L^2(\mathbb{R}^d)$. In fact, for $d \geq 3$, for any fixed positive constant M, we have

$$\begin{aligned}
\|\nabla V^\epsilon(0,x)\|_{L^2} &= \left(\int_{\mathbb{R}^d} |\nabla \triangle^{-1}(\rho_0^\epsilon - b)(x)|^2\, dx\right)^{\frac{1}{2}} \\
&= \left(\int_{\mathbb{R}^d} \left|\frac{\xi}{|\xi|^2}(\hat{\rho_0^\epsilon} - \hat{b})(\xi)\right|^2 d\xi\right)^{\frac{1}{2}} \\
&\leq \left(\int_{|\xi|\leq M} \frac{1}{|\xi|^2}|(\hat{\rho_0^\epsilon} - \hat{b})(\xi)|^2\, d\xi\right)^{\frac{1}{2}} + \frac{1}{M}\left(\int_{\mathbb{R}^d} |(\hat{\rho_0^\epsilon} - \hat{b})(\xi)|^2\, d\xi\right)^{\frac{1}{2}} \\
&\leq CM^{\frac{d-2}{2}}\|(\hat{\rho_0^\epsilon} - \hat{b})\|_{L^\infty} + \frac{1}{M}\|\rho_0^\epsilon - b\|_{L^2}, \qquad (2.5)
\end{aligned}$$

but by [19], we have

$$\|(\hat{\rho_0^\epsilon} - \hat{b})\|_{L^\infty} \leq \|\rho_0^\epsilon - b\|_{L^1}. \qquad (2.6)$$

By combining (A2) and (2.5) with (2.6), we get the estimate for $\|\nabla V^\epsilon(0,x)\|_{L^2}$ for $d \geq 3$, while for $d = 1, 2$, by (A4), we have $\int_{\mathbb{R}^d}(\rho_0^\epsilon - b)(x)\, dx = 0$, which directly implies that

$$(\hat{\rho_0^\epsilon} - \hat{b})(0) = 0;$$

then by the calculations in (2.5), we have

$$\begin{aligned}
\|\nabla V^\epsilon(0,x)\|_{L^2} &= \left(\int_{\mathbb{R}^d} \left|\frac{\xi}{|\xi|^2}\{(\hat{\rho_0^\epsilon} - \hat{b})(\xi) - (\hat{\rho_0^\epsilon} - \hat{b})(0)\}\right|^2 d\xi\right)^{\frac{1}{2}} \\
&\leq CM^{\frac{d}{2}}\|\nabla_\xi(\hat{\rho_0^\epsilon} - \hat{b})(\xi)\|_{L^\infty} + \frac{1}{M}\|\rho_0^\epsilon - b\|_{L^2} \\
&\leq CM^{\frac{d}{2}}\int_{\mathbb{R}^d} |x||\rho_0^\epsilon - b|\, dx + \frac{1}{M}\|\rho_0^\epsilon - b\|_{L^2}, \qquad (2.7)
\end{aligned}$$

which, together with (A2) and (A4), provides the estimate for $\|\nabla V^\epsilon(0,x)\|_{L^2}$.

LEMMA 2.2. *Let $s \in \mathbb{Z}^+, s > \frac{d}{2} + 2$, $(\rho_0(x), u_0(x))$ satisfy* (A3); *then there exists a positive constant $T^* > 0$ such that* (1.10) *has a unique solution $\rho(t,x) \in L^\infty([0,T], H^{s-1}(\mathbb{R}^d))$, $u(t,x) \in L^\infty([0,T], H^s(\mathbb{R}^d))$ for all $T < T^*$; moreover,*

$$\frac{d}{dt}\int_{\mathbb{R}^d} (\rho|u|^2 + |\nabla V|^2)(t,x)\, dx = 0, \qquad t < T^*, \qquad (2.8)$$

where T^* is the first time such that

$$\lim_{T\to T^*} \|\nabla u(t,x)\|_{L^\infty([0,T]\times\mathbb{R}^d)} = \infty. \qquad (2.9)$$

Proof. From the standard theory on the hyperbolic system of equations [16], we know that we can find a positive constant T such that (1.10) has a local solution (ρ, u) with $\rho(t,x) \in L^\infty([0,T], H^{s-1}(\mathbb{R}^d))$, $u(t,x) \in L^\infty([0,T], H^s(\mathbb{R}^d))$; moreover, it holds that

$$\|\rho(t,\cdot)\|_{L^\infty([0,T],H^{s-1}(\mathbb{R}^d))} + \|u(t,\cdot)\|_{L^\infty([0,T],H^s(\mathbb{R}^d))} \leq C(T, \|\rho_0\|_{H^{s-1}}, \|u_0\|_{H^s}). \qquad (2.10)$$

Hence to complete the lemma, we only need to show that if $\|\nabla u(t,x)\|_{L^\infty([0,T]\times\mathbb{R}^d)} < \infty$, then (2.8) and (2.10) hold for $t \le T$. As a convention in the rest of this section, we always assume that all the calculations are to be done for $t \le T$.

First, by multiplying (ρu) by the second equation of (1.10) and integrating over $\mathbb{R}^d$, we find

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}\rho|u|^2\,dx + \int_{\mathbb{R}^d}\rho u\nabla V\,dx = 0, \tag{2.11}$$

where we used the first equation of (1.10); then by integration by parts and (1.10) again, we have

$$\begin{aligned}\int_{\mathbb{R}^d}\rho u\nabla V\,dx &= -\int_{\mathbb{R}^d}\mathrm{div}(\rho u)V\,dx = \int_{\mathbb{R}^d}\partial_t\rho V\,dx = -\int_{\mathbb{R}^d}\partial_t\triangle V V\,dx\\ &= \frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}|\nabla V|^2\,dx.\end{aligned} \tag{2.12}$$

By summing up (2.11) and (2.12) together, we prove (2.8).

On the other hand, by (A2), (A4), and Fatou's lemma, we have

$$\begin{aligned}&\|\rho_0 - b\|_{L^1} \le \lim_{\epsilon\to 0}\|\rho_0^\epsilon - b\|_{L^1},\\ &\int_{\mathbb{R}^d}|x||\rho_0 - b|\,dx \le \overline{\lim}_{\epsilon\to 0}\int_{\mathbb{R}^d}|x||\rho_0^\epsilon - b|\,dx \quad \text{for} \quad d = 1,2.\end{aligned} \tag{2.13}$$

Thus by exactly the same proof as that of Remark 2.1, we get $\nabla V(0,x) \in L^2(\mathbb{R}^d)$. This together with (2.8) shows that

$$\int_{\mathbb{R}^d}\left(\frac{1}{2}\rho|u|^2 + |\nabla V|^2\right)(t,x)\,dx \le \int_{\mathbb{R}^d}\left(\frac{1}{2}\rho_0|u_0|^2 + |\nabla V(0,x)|^2\right)dx \le C. \tag{2.14}$$

Next we multiply (ρ, u) by (1.10) and integrate over $\mathbb{R}^d$ to obtain

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}(\rho^2 + |u|^2)(t,x)\,dx\\ &\le \frac{1}{2}\|\mathrm{div}u\|_{L^\infty}\int_{\mathbb{R}^d}(\rho^2 + |u|^2)\,dx + \|\nabla V(t,\cdot)\|_{L^2}\|u(t,\cdot)\|_{L^2}.\end{aligned} \tag{2.15}$$

On the other hand, by taking $\partial_x^{\alpha-1}$ (resp., ∂_x^α) to the first (resp., second) equation of (1.10) for $|\alpha| \le s$, and multiplying $\partial_x^{\alpha-1}\rho(t,x)$ (resp., $\partial_x^\alpha u(t,x)$) to the resulting equation and integrating over $\mathbb{R}^d$, we find

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\left(\|\partial_x^{\alpha-1}\rho(t,\cdot)\|_{L^2}^2 + \|\partial_x^\alpha u(t,\cdot)\|_{L^2}^2\right)\\ &\le \left|\int_{\mathbb{R}^d}\left\{\partial_x^{\alpha-1}\mathrm{div}(\rho u)\partial_x^{\alpha-1}\rho + \partial_x^\alpha(u\nabla u)\partial_x^\alpha u + \partial_x^\alpha\nabla V\partial_x^\alpha u\right\}dx\right|;\end{aligned} \tag{2.16}$$

however, by a Moser-type calculus inequality (see [16]), we find

$$\begin{aligned}&\left|\int_{\mathbb{R}^d}\partial_x^{\alpha-1}\mathrm{div}(\rho u)\partial_x^{\alpha-1}\rho\,dx\right|\\ &= \left|\int_{\mathbb{R}^d}u\nabla\partial_x^{\alpha-1}\rho\partial_x^{\alpha-1}\rho\,dx + \int_{\mathbb{R}^d}(\partial_x^{\alpha-1}\mathrm{div}(\rho u) - u\nabla\partial_x^{\alpha-1}\rho)\partial_x^{\alpha-1}\rho\,dx\right|\\ &\le \frac{1}{2}\|\nabla u\|_{L^\infty}\|\partial_x^{\alpha-1}\rho\|_{L^2}^2 + \{\|\nabla u\|_{L^\infty}\|\partial_x^{|\alpha|-1}\rho\|_{L^2} + \|\rho\|_{L^\infty}\|\partial_x^{|\alpha|}u\|_{L^2}\}\|\partial_x^{\alpha-1}\rho\|_{L^2}.\end{aligned} \tag{2.17}$$

Exactly as in the proof of (2.17), we have

$$\left|\int_{\mathbb{R}^d} \partial_x^\alpha(u\nabla u)\partial_x^\alpha u\,dx\right| \le \frac{1}{2}\|\nabla u\|_{L^\infty}\|\partial_x^\alpha u\|_{L^2}^2 + 2\|\nabla u\|_{L^\infty}\|\partial_x^{|\alpha|}u\|_{L^2}\|\partial_x^\alpha u\|_{L^2}, \tag{2.18}$$

and, trivially,

$$\begin{aligned}\left|\int_{\mathbb{R}^d} \partial_x^\alpha \nabla V \partial_x^\alpha u\,dx\right| &= \left|\int_{\mathbb{R}^d} \partial_x^\alpha \nabla\triangle^{-1}(\rho-b)\partial_x^\alpha u\,dx\right| \\ &\le (C + \|\partial_x^{\alpha-1}\rho\|_{L^2})\|\partial_x^\alpha u\|_{L^2}.\end{aligned} \tag{2.19}$$

By summing up (2.15)–(2.19), and doing summation for $|\alpha| \le s$, we finally get

$$\frac{d}{dt}\left(\|\rho(t,\cdot)\|_{H^{s-1}}^2 + \|u(t,\cdot)\|_{H^s}^2\right) \le C(1+\|\nabla u\|_{L^\infty})\left(\|\rho(t,\cdot)\|_{H^{s-1}}^2 + \|u(t,\cdot)\|_{H^s}^2\right). \tag{2.20}$$

Equation (2.20) together with the Gronwall inequality implies that if

$$\|\nabla u(t,\cdot)\|_{L^\infty([0,T]\times\mathbb{R}^d)} < \infty,$$

then (2.10) holds. This completes the proof of the lemma. □

3. Proof of Theorem 1.1. With the fundamental lemmas, Lemmas 2.1 and 2.2, motivated by [1], for $t < T^*$, let us define

$$H_u^\epsilon(t) = \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\xi - u(t,x)|^2 f^\epsilon(t,x,\xi)\,d\xi\,dx + \frac{1}{2}\int_{\mathbb{R}^d} |\nabla\triangle^{-1}(\rho^\epsilon - \rho)|^2\,dx, \tag{3.1}$$

where (ρ, u) is the unique local smooth solution of (1.10), and T^* is the positive constant determined by Lemma 2.2. In the following, we are going to show that $\lim_{\epsilon\to 0} H_u^\epsilon(t) = 0$ for $0 \le t < T^*$, which directly implies that the Wigner measure $f(t,x,\xi)$, which is the weak limit of $f^\epsilon(t,x,\xi)$ in some sense (see [14] or [7]), equals $\rho(t,x)\delta(\xi - u(t,x))$. The idea of the proof is somewhat similar to that used in [22] and [24], where we needed to control the variance of the L^p Young measures associated with the approximate solution sequences to the problems there. So now we need to derive an evolution equation for $H_u^\epsilon(t)$.

To begin, let us calculate the first two moment equations of (1.6).

LEMMA 3.1. *Let* $\rho^\epsilon(t,x) =: |\psi^\epsilon(t,x)|^2$, $J^\epsilon(t,x) =: \epsilon Im(\overline{\psi^\epsilon}\nabla\psi^\epsilon)(t,x)$; *then the following hold:*

(1)

$$\partial_t\rho^\epsilon + \mathrm{div}J^\epsilon = 0, \tag{3.2}$$

$$\partial_t J^\epsilon + \nabla_x : \int_{\mathbb{R}^d} \xi\otimes\xi f^\epsilon\,d\xi + \nabla V^\epsilon\rho^\epsilon = 0; \tag{3.3}$$

(2)

$$\frac{d}{dt}\left\{\frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f^\epsilon|\xi|^2\,dx\,d\xi + \int_{\mathbb{R}^d}|\nabla V^\epsilon|^2\,dx\right\} = 0. \tag{3.4}$$

Proof. First, notice by [14] (or [7]) that we have $\int_{\mathbb{R}^d} f^\epsilon(t,x,\xi)\,d\xi = \rho^\epsilon(t,x)$, and by Lemma 2.1, for any fixed $\epsilon>0, t, x$, $D_y\{\psi^\epsilon(t,x+\frac{\epsilon y}{2})\overline{\psi^\epsilon(t,x-\frac{\epsilon y}{2})}\} \in (L^1_y\cap H^s_y)(\mathbb{R}^d)$ with $s>\frac{d}{2}+1$. Thus by Remark 3.1, we have

$$\begin{aligned}\int_{\mathbb{R}^d}\xi f^\epsilon(t,x,\xi)\,d\xi &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\xi y}\xi\psi^\epsilon\left(t,x+\frac{\epsilon y}{2}\right)\overline{\psi^\epsilon\left(t,x-\frac{\epsilon y}{2}\right)}\,dy\,d\xi\\ &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\xi y}D_y\left\{\psi^\epsilon\left(t,x+\frac{\epsilon y}{2}\right)\overline{\psi^\epsilon\left(t,x-\frac{\epsilon y}{2}\right)}\right\}\,dy\,d\xi\\ &= D_y\left\{\psi^\epsilon\left(t,x+\frac{\epsilon y}{2}\right)\overline{\psi^\epsilon\left(t,x-\frac{\epsilon y}{2}\right)}\right\}\Big|_{y=0} = J^\epsilon(t,x). \end{aligned}\tag{3.5}$$

And by (1.7), we have

(3.6)

$$\begin{aligned}&\int_{\mathbb{R}^d}\xi^\alpha\theta[V^\epsilon]f^\epsilon(t,x,\xi)\,d\xi\\ &= \frac{i}{(2\pi)^d}\int_{\mathbb{R}^d} f^\epsilon(t,x,\eta)\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i(\xi-\eta)y}\xi^\alpha\frac{V^\epsilon(t,x+\frac{\epsilon y}{2})-V^\epsilon(t,x-\frac{\epsilon y}{2})}{\epsilon}\,dy\,d\xi\,d\eta\\ &= \frac{i}{(2\pi)^d}\int_{\mathbb{R}^d} f^\epsilon(t,x,\eta)\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\xi y}D_y^\alpha\left(\frac{V^\epsilon(t,x+\frac{\epsilon y}{2})-V^\epsilon(t,x-\frac{\epsilon y}{2})}{\epsilon}e^{-i\eta y}\right)\,dy\,d\xi\,d\eta.\end{aligned}$$

In particular, by taking $|\alpha|=0$ and $|\alpha|=1$ in (3.6), Remark 3.1 directly implies that

$$\begin{aligned}&\int_{\mathbb{R}^d}\theta[V^\epsilon]f^\epsilon(t,x,\xi)\,d\xi = 0,\\ &\int_{\mathbb{R}^d}\xi\theta[V^\epsilon]f^\epsilon(t,x,\xi)\,d\xi = \int_{\mathbb{R}^d} f^\epsilon(t,x,\eta)\nabla V^\epsilon(t,x)\,d\eta = (\rho^\epsilon\nabla V^\epsilon)(t,x).\end{aligned}\tag{3.7}$$

With (3.5)–(3.7), by integrating (1.6) over $\mathbb{R}^d$ with respect to ξ, we get (3.2), and by multiplying (1.6) by ξ and integrating over $\mathbb{R}^d$ with respect to ξ again, we find that (3.3) holds.

On the other hand, exactly as in the proof of (3.5), we have

$$\begin{aligned}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|\xi|^2 f^\epsilon(t,x,\xi)\,d\xi\,dx &= -\frac{\epsilon^2}{4}\int_{\mathbb{R}^d}\{\overline{\psi^\epsilon}\triangle\psi^\epsilon - 2\nabla\psi^\epsilon\nabla\overline{\psi^\epsilon}+\psi^\epsilon\triangle\overline{\psi^\epsilon}\}\,dx\\ &= \epsilon^2\int_{\mathbb{R}^d}|\nabla\psi^\epsilon|^2\,dx.\end{aligned}\tag{3.8}$$

Equation (2.2) together with (3.8) shows that (3.4) holds. This completes the proof of Lemma 3.1. □

Remark 3.1. Let $s>\frac{d}{2}$, $f(y)\in(L^1\cap H^s)(\mathbb{R}^d)$; then we claim that

$$\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\xi y}f(y)\,dy\,d\xi = f(0).\tag{3.9}$$

In fact, as $f(y)\in L^1(\mathbb{R}^d)$, by [19], $\hat{f}(\xi) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\xi y}f(y)\,dy$. On the other hand, by the fact that $f(y)\in H^s(\mathbb{R}^d)$ for $s>\frac{d}{2}$, we have

$$\hat{f}(\xi) = \frac{1}{(1+|\xi|^2)^{\frac{s}{2}}}\widehat{\psi^s(D)f}(\xi)\in L^1(\mathbb{R}^d),\tag{3.10}$$

where $\psi^s(D)$ is the pseudodifferential operator (see [21]) with symbol $(1+|\xi|^2)^{\frac{s}{2}}$. Hence by (3.10) and [19] again, we have

$$f(y) = \int_{\mathbb{R}^d} e^{i\xi y}\hat{f}(\xi)\,d\xi. \tag{3.11}$$

In particular, by taking $y=0$ in (3.11), we prove (3.9).

Next let us derive the evolution equation for $H_u^\epsilon(t)$.

LEMMA 3.2. *For $H_u^\epsilon(t)$ defined in* (3.1), *there holds*

$$\begin{aligned}\frac{d}{dt}H_u^\epsilon(t) = &-\int_{\mathbb{R}^d} Du : \int_{\mathbb{R}^d}(\xi-u)\otimes(\xi-u)f^\epsilon\,d\xi\,dx - \frac{1}{2}\int_{\mathbb{R}^d}\mathrm{div}u|\nabla\triangle^{-1}(\rho-\rho^\epsilon)|^2\,dx\\ &+\int_{\mathbb{R}^d} Du : \left(\nabla\triangle^{-1}(\rho^\epsilon-\rho)\otimes\nabla\triangle^{-1}(\rho^\epsilon-\rho)\right)\,dx.\end{aligned} \tag{3.12}$$

Proof. First, by (3.4), we find

$$\begin{aligned}\frac{d}{dt}H_u^\epsilon(t) = &\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}|u|^2\rho^\epsilon\,dx - \frac{d}{dt}\int_{\mathbb{R}^d}uJ^\epsilon\,dx\\ &+\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}|\nabla\triangle^{-1}(\rho-b)|^2\,dx - \frac{d}{dt}\int_{\mathbb{R}^d}\nabla\triangle^{-1}(\rho^\epsilon-b)\nabla\triangle^{-1}(\rho-b)\,dx.\end{aligned} \tag{3.13}$$

In the following, we are going to calculate the above four terms separately. First, by (3.2), we have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}|u|^2\rho^\epsilon\,dx &= \int_{\mathbb{R}^d}\left\{u\partial_t u\rho^\epsilon + \frac{1}{2}|u|^2\partial_t\rho^\epsilon\right\}dx\\ &= \int_{\mathbb{R}^d}\left\{\partial_t u\rho^\epsilon u - \frac{1}{2}|u|^2\mathrm{div}J^\epsilon\right\}dx\\ &= \int_{\mathbb{R}^d}\left\{\partial_t u\rho^\epsilon u + \frac{1}{2}J^\epsilon\nabla|u|^2\right\}dx,\end{aligned} \tag{3.14}$$

and by (3.3), we have

$$\begin{aligned}-\frac{d}{dt}\int_{\mathbb{R}^d}J^\epsilon u\,dx &= \int_{\mathbb{R}^d}\{-\partial_t J^\epsilon u - J^\epsilon\partial_t u\}\,dx\\ &= \int_{\mathbb{R}^d}\left\{\left(\nabla_x : \int_{\mathbb{R}^d}\xi\otimes\xi f^\epsilon\,d\xi + \nabla V^\epsilon\rho^\epsilon\right)u - J^\epsilon\partial_t u\right\}dx\\ &= -\int_{\mathbb{R}^d}Du : \int_{\mathbb{R}^d}\xi\otimes\xi f^\epsilon\,d\xi\,dx + \int_{\mathbb{R}^d}\nabla V^\epsilon\rho^\epsilon u\,dx - \int_{\mathbb{R}^d}J^\epsilon\partial_t u\,dx,\end{aligned} \tag{3.15}$$

while by (1.10), we have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}|\nabla\triangle^{-1}(\rho-b)|^2\,dx &= \int_{\mathbb{R}^d}\nabla\triangle^{-1}(\rho-b)\nabla\triangle^{-1}\partial_t\rho\,dx\\ &= -\int_{\mathbb{R}^d}\nabla\triangle^{-1}(\rho-b)\nabla\triangle^{-1}\mathrm{div}(\rho u)\,dx\\ &= \int_{\mathbb{R}^d}\triangle^{-1}(\rho-b)\mathrm{div}(\rho u)\,dx\\ &= -\int_{\mathbb{R}^d}\rho u\nabla\triangle^{-1}(\rho-b)\,dx.\end{aligned} \tag{3.16}$$

Then by (1.10), (3.2), and a calculation similar to that in (3.16), we have

$$
\begin{aligned}
&-\frac{d}{dt}\int_{\mathbb{R}^d}\nabla\triangle^{-1}(\rho^\epsilon-b)\nabla\triangle^{-1}(\rho-b)\,dx\\
&=-\int_{\mathbb{R}^d}\nabla\triangle^{-1}\partial_t\rho^\epsilon\nabla\triangle^{-1}(\rho-b)\,dx-\int_{\mathbb{R}^d}\nabla\triangle^{-1}(\rho^\epsilon-b)\nabla\triangle^{-1}\partial_t\rho\,dx\\
&=\int_{\mathbb{R}^d}J^\epsilon\nabla\triangle^{-1}(\rho-b)\,dx+\int_{\mathbb{R}^d}\rho u\nabla\triangle^{-1}(\rho^\epsilon-b)\,dx. \qquad (3.17)
\end{aligned}
$$

By summing (3.13)–(3.17), we get

$$
\begin{aligned}
\frac{d}{dt}H_u^\epsilon(t)&=\int_{\mathbb{R}^d}\Big\{\big(\partial_t u+u\nabla u-\nabla\triangle^{-1}(\rho-b)\big)\rho^\epsilon u-u\nabla u\rho^\epsilon u+\nabla\triangle^{-1}(\rho-b)\rho^\epsilon u\\
&\quad+\frac{1}{2}J^\epsilon\nabla|u|^2-Du:\int_{\mathbb{R}^d}\xi\otimes\xi f^\epsilon\,d\xi+\nabla V^\epsilon\rho^\epsilon u\\
&\quad-J^\epsilon\big(\partial_t u+u\nabla u-\nabla\triangle^{-1}(\rho-b)\big)+u\nabla uJ^\epsilon+\rho u\nabla\triangle^{-1}(\rho^\epsilon-\rho)\Big\}\\
&=\int_{\mathbb{R}^d}\Big\{-u\nabla u\rho^\epsilon u-Du:\int_{\mathbb{R}^d}\xi\otimes\xi f^\epsilon\,d\xi+\frac{1}{2}J^\epsilon\nabla|u|^2+u\nabla uJ^\epsilon\\
&\quad+\nabla\triangle^{-1}(\rho-b)\rho^\epsilon u+\nabla V^\epsilon\rho^\epsilon u+\rho u\nabla\triangle^{-1}(\rho^\epsilon-\rho)\Big\}\,dx\\
&=-\int_{\mathbb{R}^d}Du:\int_{\mathbb{R}^d}(\xi-u)\otimes(\xi-u)f^\epsilon\,d\xi\,dx \qquad (3.18)\\
&\quad+\int_{\mathbb{R}^d}\big\{\nabla\triangle^{-1}(\rho-b)\rho^\epsilon u-\nabla\triangle^{-1}(\rho^\epsilon-b)\rho^\epsilon u+\rho u\nabla\triangle^{-1}(\rho^\epsilon-\rho)\big\}\,dx,
\end{aligned}
$$

where in the second step of the above derivation, we used (1.10).

On the other hand, notice that for any $a(x)\in C^1(\mathbb{R}^d)$, it holds that

$$
\nabla:(\nabla a\otimes\nabla a)-\frac{1}{2}\nabla|\nabla a|^2=\nabla a\triangle a,
$$

and we get

$$
\begin{aligned}
&-\int_{\mathbb{R}^d}\nabla\triangle^{-1}(\rho^\epsilon-b)\rho^\epsilon u\,dx+\int_{\mathbb{R}^d}\nabla\triangle^{-1}(\rho^\epsilon-b)bu\,dx\\
&=-\int_{\mathbb{R}^d}\nabla:\big(\nabla\triangle^{-1}(\rho^\epsilon-b)\otimes\nabla\triangle^{-1}(\rho^\epsilon-b)\big)\,u\,dx\\
&+\frac{1}{2}\int_{\mathbb{R}^d}\nabla\big(|\nabla\triangle^{-1}(\rho^\epsilon-b)|^2\big)\,u\,dx\\
&=\int_{\mathbb{R}^d}Du:\big(\nabla\triangle^{-1}(\rho^\epsilon-b)\otimes\nabla\triangle^{-1}(\rho^\epsilon-b)\big)\,dx-\frac{1}{2}\int_{\mathbb{R}^d}\mathrm{div}u|\nabla\triangle^{-1}(\rho^\epsilon-b)|^2\,dx.
\end{aligned}
\qquad (3.19)
$$

Exactly as in the proof of (3.19), we have

$$
\begin{aligned}
&-\int_{\mathbb{R}^d}\nabla\triangle^{-1}(\rho-b)\rho u\,dx+\int_{\mathbb{R}^d}\nabla\triangle^{-1}(\rho-b)bu\,dx\\
&=\int_{\mathbb{R}^d}Du:\big(\nabla\triangle^{-1}(\rho-b)\otimes\nabla\triangle^{-1}(\rho-b)\big)\,dx-\frac{1}{2}\int_{\mathbb{R}^d}\mathrm{div}u|\nabla\triangle^{-1}(\rho-b)|^2\,dx.
\end{aligned}
\qquad (3.20)
$$

Moreover, for any two $C^1(\mathbb{R}^d)$ functions $a(x), b(x)$, it holds that

$$\nabla : (\nabla a \otimes \nabla b) + \nabla : (\nabla b \otimes \nabla a) - \nabla(\nabla a \cdot \nabla b) = \nabla a \triangle b + \nabla b \triangle a,$$

and we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left\{\nabla\triangle^{-1}(\rho - b)\rho^\epsilon u + \nabla\triangle^{-1}(\rho^\epsilon - b)\rho u\right\} dx \\
&- \int_{\mathbb{R}^d} \left\{\nabla\triangle^{-1}(\rho - b) + \nabla\triangle^{-1}(\rho^\epsilon - b)\right\} bu\, dx \\
&= \int_{\mathbb{R}^d} \{\nabla : \left(\nabla\triangle^{-1}(\rho - b) \otimes \nabla\triangle^{-1}(\rho^\epsilon - b)\right) \\
&+ \nabla : \left(\nabla\triangle^{-1}(\rho^\epsilon - b) \otimes \nabla\triangle^{-1}(\rho - b)\right)\} u\, dx \\
&- \int_{\mathbb{R}^d} \nabla\left(\nabla\triangle^{-1}(\rho^\epsilon - b) \cdot \nabla\triangle^{-1}(\rho - b)\right) u\, dx \qquad (3.21) \\
&= -\int_{\mathbb{R}^d} \{\left(\nabla\triangle^{-1}(\rho - b) \otimes \nabla\triangle^{-1}(\rho^\epsilon - b)\right) \\
&\qquad + \left(\nabla\triangle^{-1}(\rho^\epsilon - b) \otimes \nabla\triangle^{-1}(\rho - b)\right)\} : Du\, dx \\
&+ \int_{\mathbb{R}^d} \mathrm{div} u \nabla\triangle^{-1}(\rho^\epsilon - b) \cdot \nabla\triangle^{-1}(\rho - b)\, dx.
\end{aligned}$$

Then by summing up (3.19)–(3.21) and an appropriate rearrangement, we have

(3.22)

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left\{\nabla\triangle^{-1}(\rho - b)\rho^\epsilon u - \nabla\triangle^{-1}(\rho^\epsilon - b)\rho^\epsilon u + \rho u \nabla\triangle^{-1}(\rho^\epsilon - \rho)\right\} dx \\
&= \int_{\mathbb{R}^d} Du : \left(\nabla\triangle^{-1}(\rho^\epsilon - \rho) \otimes \nabla\triangle^{-1}(\rho^\epsilon - \rho)\right) dx - \frac{1}{2}\int_{\mathbb{R}^d} \mathrm{div} u |\nabla\triangle^{-1}(\rho^\epsilon - \rho)|^2\, dx.
\end{aligned}$$

From (3.18) and (3.22), we get (3.12). □

With (3.12), now we can prove that the Wigner measure of $\psi^\epsilon(t,x)$ is $\rho(t,x)\delta(\xi - u(t,x))$ before the formation of singularities in (1.10).

LEMMA 3.3. *Let $f^\epsilon(t,x,\xi)$ be the Wigner transformation of $\psi^\epsilon(t,x)$; then for any fixed $t < T^*$, it holds that*

$$f^\epsilon(t,x,\xi) \rightharpoonup f(t,x,\xi) =: \rho(t,x)\delta(\xi - u(t,x)) \quad \text{in} \quad \mathcal{A}'(\mathbb{R}^{2d}) \quad \text{as} \quad \epsilon \to 0. \tag{3.23}$$

Proof. Compared with the situation in [1], a new difficulty arising here is the sign changing of the Wigner transformation $f^\epsilon(t,x,\xi)$. However, by the following elementary calculations, we fortunately find that $H_u^\epsilon(t)$ keeps its positive sign. In fact, by (3.5) and (3.8), we have

$$\begin{aligned}
&\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f^\epsilon(t,x,\xi)|\xi - u(t,x)|^2\, dx\, d\xi \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f^\epsilon(t,x,\xi)|\xi|^2\, d\xi\, dx + \int_{\mathbb{R}^d} |u|^2|\psi^\epsilon|^2\, dx - 2\int_{\mathbb{R}^d} uJ^\epsilon\, dx \\
&= \int_{\mathbb{R}^d} \left\{\epsilon^2|\nabla\psi^\epsilon|^2 - 2\epsilon\mathrm{Im}(\overline{\psi^\epsilon}\nabla\psi^\epsilon)u + |u|^2|\psi^\epsilon|^2\right\} dx \\
&= \int_{\mathbb{R}^d} |(u - \epsilon D)\psi^\epsilon|^2\, dx, \qquad (3.24)
\end{aligned}$$

with $D = \frac{1}{i}\nabla_x$, while by a calculation similar to that of (3.8), we have

$$\int_{\mathbb{R}^d} \xi_i\xi_j f^\epsilon \, d\xi = -\frac{\epsilon^2}{4}\left(\partial_i\partial_j\psi^\epsilon\overline{\psi^\epsilon} - \partial_i\psi^\epsilon\partial_j\overline{\psi^\epsilon} - \partial_i\overline{\psi^\epsilon}\partial_j\psi^\epsilon + \psi^\epsilon\partial_i\partial_j\overline{\psi^\epsilon}\right),$$

from which we get

(3.25)
$$\begin{aligned}
&\left|\int_{\mathbb{R}^d} \partial_i u_j \int_{\mathbb{R}^d} (\xi-u)_i(\xi-u)_j f^\epsilon \, d\xi\, dx\right| \\
&= \left|\int_{\mathbb{R}^d} \left\{\partial_i u_j \left(\epsilon^2 \mathrm{Re}(\partial_i\psi^\epsilon\partial_j\overline{\psi^\epsilon}) - \epsilon u_i \mathrm{Im}(\overline{\psi^\epsilon}\partial_j\psi^\epsilon) - \epsilon u_j \mathrm{Im}(\overline{\psi^\epsilon}\partial_i\psi^\epsilon) + u_iu_j|\psi^\epsilon|^2\right)\right. \\
&\left. + \frac{\epsilon^2}{2}\partial_i^2 u_j \mathrm{Re}(\overline{\psi^\epsilon}\partial_j\psi^\epsilon)\right\} dx\Big|.
\end{aligned}$$

Then by rearranging the above, we immediately get

(3.26)
$$\begin{aligned}
&\left|\int_{\mathbb{R}^d} Du : \int_{\mathbb{R}^d} (\xi-u)\otimes(\xi-u) f^\epsilon \, d\xi\, dx\right| \\
&\le \sum_{i,j=1}^d \left|\int_{\mathbb{R}^d} \partial_i u_j \int_{\mathbb{R}^d} (\xi-u)_i(\xi-u)_j f^\epsilon \, d\xi\, dx\right| \\
&= \left|\sum_{i,j=1}^d \left\{\int_{\mathbb{R}^d} \partial_i u_j \mathrm{Re}\left((u_i - \epsilon D_i)\psi^\epsilon \overline{(u_j-\epsilon D_j)\psi^\epsilon}\right) dx + \frac{\epsilon^2}{2}\int_{\mathbb{R}^d} \partial_i^2 u_j \mathrm{Re}(\overline{\psi^\epsilon}\partial_j\psi^\epsilon)\, dx\right\}\right| \\
&\le C\left\{\|\nabla u\|_{L^\infty}\|(u-\epsilon D)\psi^\epsilon\|_{L^2}^2 + \epsilon\|\nabla_x^2 u\|_{L^\infty}\|\psi^\epsilon\|_{L^2}\|\epsilon\nabla\psi^\epsilon\|_{L^2}\right\}.
\end{aligned}$$

By summing up Lemma 2.1, (3.12), (3.24), and (3.26), we find, for any $0 \le t < T^*$, it holds that

$$\frac{d}{dt}H_u^\epsilon(t) \le C\|\nabla u(t,\cdot)\|_{L^\infty} H_u^\epsilon(t) + C\epsilon. \tag{3.27}$$

And by (A2), (A3) in the introduction and (3.24), we have

$$\begin{aligned}
H_u^\epsilon(t)|_{t=0} &= \frac{1}{2}\int_{\mathbb{R}^d} |(u_0-\epsilon D)\psi_0^\epsilon|^2\, dx + \int_{\mathbb{R}^d} |\nabla\triangle^{-1}(\rho_0^\epsilon - \rho_0)|^2\, dx \\
&\le \int_{\mathbb{R}^d} \rho_0^\epsilon |u_0 - \nabla S^\epsilon|^2\, dx + \epsilon\int_{\mathbb{R}^d} |\nabla\sqrt{\rho_0^\epsilon}|^2\, dx + \int_{\mathbb{R}^d} |\nabla\triangle^{-1}(\rho_0^\epsilon-\rho_0)|^2\, dx \\
&= o(1). \qquad (3.28)
\end{aligned}$$

By summing up (3.27), (3.28), and Gronwall's inequality, we find

$$\lim_{\epsilon\to 0} H_u^\epsilon(t) = 0 \qquad \text{for all} \qquad t < T^*. \tag{3.29}$$

Now let us take a smooth cut-off function $\chi(x,\xi) \in C_c^\infty(\mathbb{R}^d\times\mathbb{R}^d)$ with

$$\chi(x,\xi) = 1 \quad \text{for} \quad |x|+|\xi| \le 1, \quad \mathrm{supp}\chi \subset B(0,2); \tag{3.30}$$

then by Proposition 1.1 of [7], we get

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\chi^2\left(\frac{x}{R},\frac{\xi}{R}\right)|\xi-u(t,x)|^2 f^\epsilon(t,x,\xi)\,dx\,d\xi$$

$$= \int_{\mathbb{R}^d}\left|\chi\left(\frac{x}{R},\frac{\epsilon D}{R}\right)[(u(t,x)-\epsilon D)\psi^\epsilon(t,x)]\right|^2 dx + r_\epsilon, \tag{3.31}$$

where $|r_\epsilon| \le \epsilon C(\chi,u)\|\psi^\epsilon\|_{L^2}^2$, and by Lemma 0.5 D of [21], we have

$$\left\|\chi\left(\frac{x}{R},\frac{\epsilon D}{R}\right)[(u(t,x)-\epsilon D)\psi^\epsilon(t,x)]\right\|_{L^2} \le C\|(u(t,x)-\epsilon D)\psi^\epsilon(t,x)\|_{L^2}. \tag{3.32}$$

On the other hand, by [14] (or [7]), for any fixed $t < T^*$ there is a subsequence $\{f^\epsilon(t,x,\xi)\}$ (we still denote it by $\{f^\epsilon\}$ for convenience) and a nonnegative Radon measure $f(t,x,\xi)$, which is called the Wigner measure of $\{\psi^\epsilon(t,x)\}$, such that

$$f^\epsilon(t,x,\xi) \rightharpoonup f(t,x,\xi) \quad \text{in} \quad \mathcal{A}'(\mathbb{R}^{2d}) \quad \text{as} \quad \epsilon\to 0. \tag{3.33}$$

Furthermore, by (A2), Lemma 2.1, and Remark 2.1, we have

$$\epsilon\|\nabla_x\psi^\epsilon(t,x)\|_{L^2} \le C, \tag{3.34}$$

which together with [14, Theorem III.1, Remark III.13] (or [7]) shows that

$$\rho^\epsilon(t,x) = \int_{\mathbb{R}^d} f^\epsilon(t,x,\xi)\,d\xi \rightharpoonup \int_{\mathbb{R}^d} f(t,x,d\xi) \quad \text{in} \quad \mathcal{M}^+(\mathbb{R}^d), \tag{3.35}$$

while by (3.29), we have $\rho^\epsilon(t,x)\to\rho(t,x)$ in $L^\infty([0,T],H^{-1}_{\mathrm{loc}}(\mathbb{R}^d))$ for any $T<T^*$, which together with (3.35) shows that

$$\int_{\mathbb{R}^d} f(t,x,d\xi) = \rho(t,x) \tag{3.36}$$

for any fixed $t<T^*$.

In particular, by (3.32)–(3.33) for the cut-off function $\chi(x,\xi)$ chosen as above, we have

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\chi^2\left(\frac{x}{R},\frac{\xi}{R}\right)|\xi-u(t,x)|^2 f(t,dx,d\xi) \tag{3.37}$$

$$= \lim_{\epsilon\to 0}\int_{\mathbb{R}^d}\chi^2\left(\frac{x}{R},\frac{\xi}{R}\right)|\xi-u(t,x)|^2 f^\epsilon(t,x,\xi)\,dx\,d\xi$$

$$\le C\lim_{\epsilon\to 0}\|(u(t,x)-\epsilon D)\psi^\epsilon(t,x)\|_{L^2}^2 \le C\lim_{\epsilon\to 0} H_u^\epsilon(t) = 0. \tag{3.38}$$

Notice that $f(t,\cdot,\cdot)\in\mathcal{M}^+(\mathbb{R}^{2d})$, and Fatou's lemma implies that

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|\xi-u(t,x)|^2 f(t,dx,d\xi) = 0. \tag{3.39}$$

Thus for any test function $\phi(x,\xi)\in C_c^\infty(\mathbb{R}^d\times\mathbb{R}^d)$, it holds that

$$\left|\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi(x,\xi)\,df(t,x,\xi) - \int_{\mathbb{R}^d}\phi(x,u(t,x))\int_{\mathbb{R}^d} df(t,x,d\xi)\right|$$

$$\le \sup_{x,\xi\in\mathbb{R}^d}|\nabla_\xi\phi|\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|\xi-u(t,x)|\,df(t,x,\xi)$$

$$\le \sup_{x,\xi\in\mathbb{R}^d}|\nabla_\xi\phi|\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} df(t,x,\xi)\right)^{\frac12}\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|\xi-u(t,x)|^2\,df(t,x,\xi)\right)^{\frac12},$$

$$= 0,$$

which together with (3.36) implies that

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi(x,\xi)\,df(t,x,\xi)=\int_{\mathbb{R}^d}\phi(x,u(t,x))\int_{\mathbb{R}^d}df(t,x,d\xi)$$
$$=\int_{\mathbb{R}^d}\phi(x,u(t,x))\rho(t,x)\,dx=\left(\phi(x,\xi),\quad \rho(t,x)\delta(\xi-u(t,x))\right),$$

that is,

$$f(t,x,\xi)=\rho(t,x)\delta(\xi-u(t,x)). \tag{3.40}$$

But as (ρ,u) is the unique local solution of (1.10), the above argument in fact implies that we do not need to take the subsequence of $\{f^\epsilon(t,x,\xi)\}$; that is, (3.23) holds for any fixed $t<T^*$. This completes the proof of the lemma. □

LEMMA 3.4. *Let $f(t,x,\xi)$ be the measure defined in* (3.23); *then for every fixed $t<T^*$, it holds that*

$$\epsilon\mathrm{Im}(\overline{\psi^\epsilon}\nabla\psi^\epsilon)(t,x)\rightharpoonup\int_{\mathbb{R}^d}\xi f(t,x,d\xi)=(\rho u)(t,x)\quad in\quad \mathcal{M}(\mathbb{R}^d)\quad as\quad \epsilon\to 0. \tag{3.41}$$

Proof. For a clear presentation, as in [7], let us define

$$W^\epsilon(f,g)=\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}e^{-iz\xi}f\left(x+\frac{\epsilon z}{2}\right)\overline{g\left(x-\frac{\epsilon z}{2}\right)}\,dz. \tag{3.42}$$

Then for any test function $\phi(x)\in C_c^\infty(\mathbb{R}^d)$ and any cut-off function $\chi(\xi)\in C_c^\infty(\mathbb{R}^d)$, with

$$\chi(\xi)=1\quad\text{for}\quad|\xi|\le 1,\quad \mathrm{supp}\chi(\cdot)\subset B(0,2), \tag{3.43}$$

we claim that for $f(x),g(x)\in H^s(\mathbb{R}^d)$ for $s>\frac{d}{2}$, it holds that

$$\left(W^\epsilon(f,g),\phi(x)\left(1-\chi\left(\frac{\xi}{R}\right)\right)\right)=\left(f\phi,\left(1-\chi\left(\frac{\epsilon D}{R}\right)\right)g\right)+r_\epsilon, \tag{3.44}$$

where $\chi(\frac{\epsilon D}{R})$ is the pseudodifferential operator with symbol $\chi(\frac{\epsilon\xi}{R})$ (see [21]), (a,b) denotes the (complex) L^2 inner product, and

$$|r_\epsilon|\le\frac{C\epsilon}{R}\|\nabla\phi\|_{L^\infty}\|f\|_{L^2}\|g\|_{L^2}. \tag{3.45}$$

In fact, by (3.42) and the change of variables that $x'=x+\frac{\epsilon z}{2}$, we have

(3.46)

$$\left(W^\epsilon(f,g),\phi(x)\left(1-\chi\left(\frac{\xi}{R}\right)\right)\right)$$
$$=\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}e^{-iz\xi}f(x')\overline{g(x'-\epsilon z)}\phi\left(x'-\frac{\epsilon z}{2}\right)\left(1-\chi\left(\frac{\xi}{R}\right)\right)\,dz\,dx'\,d\xi$$
$$=\int_{\mathbb{R}^d}f(x')\phi(x')\frac{1}{(2\pi)^d}\overline{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}e^{iz\xi}\left(1-\chi\left(\frac{\xi}{R}\right)\right)g(x'-\epsilon z)\,dz\,d\xi\,dx'}$$
$$-\frac{1}{2(2\pi)^d}\int_{\mathbb{R}^d}f(x')\int_{-1}^0\epsilon z\nabla\phi\left(x'+\frac{\epsilon\theta z}{2}\right)\,d\theta\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}e^{-iz\xi}\left(1-\chi\left(\frac{\xi}{R}\right)\right)\overline{g(x'-\epsilon z)}\,dz\,d\xi\,dx'$$
$$=\left(f\phi,\left(1-\chi\left(\frac{\epsilon D}{R}\right)\right)g\right)+r_\epsilon.$$

But by Remark 3.1 again, we have

$$\begin{aligned}
&\frac{1}{2(2\pi)^d}\int_{\mathbb{R}^d} f(x')\int_{-1}^{0}\epsilon z\nabla\phi\left(x'+\frac{\epsilon\theta z}{2}\right)d\theta\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-iz\xi}\overline{g(x'-\epsilon z)}\,dz\,d\xi\,dx'\\
&=\frac{\epsilon}{2}\int_{\mathbb{R}^d}\int_{-1}^{0} f(x')\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-iz\xi}z\nabla\phi\left(x'+\frac{\epsilon\theta z}{2}\right)\overline{g(x'-\epsilon z)}\,dz\,d\xi\,d\theta\,dx'\\
&=0. \qquad (3.47)
\end{aligned}$$

And for $\chi(\xi)\in C_c^\infty(\mathbb{R}^d)$, we can use integration by parts to obtain

(3.48)

$$\begin{aligned}
|r_\epsilon| &= \left|\frac{\epsilon}{2(2\pi)^d}\int_{\mathbb{R}^d} f(x')\int_{-1}^{0}\nabla\phi\left(x'+\frac{\epsilon\theta z}{2}\right)d\theta\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-iz\xi}\frac{1}{R}\chi'\left(\frac{\xi}{R}\right)\overline{g(x'-\epsilon z)}\,dz\,d\xi\,dx'\right|\\
&=\left|\frac{\epsilon}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(x')\int_{-1}^{0}\nabla\phi\left(x'+\frac{\epsilon\theta z}{2}\right)d\theta\overline{g(x'-\epsilon z)}R^{d-1}\hat{\chi}'(Rz)\,dx'\,dz\right|\\
&\le\frac{\epsilon}{2}R^{d-1}\int_{\mathbb{R}^d}|\hat{\chi}'(Rz)|\,dz\|f\|_{L^2}\|g\|_{L^2}\|\nabla\phi\|_{L^\infty}\\
&\le\frac{C\epsilon}{R}\|\nabla\phi\|_{L^\infty}\|f\|_{L^2}\|g\|_{L^2}.
\end{aligned}$$

By summing up (3.46)–(3.48), we complete the proof of (3.44) and (3.45). And if we use the change of variables $x'=x-\frac{\epsilon z}{2}$ in (3.46), then by the same proof as the above, we get

$$\left(W^\epsilon(f,g),\phi(x)\left(1-\chi\left(\frac{\xi}{R}\right)\right)\right)=\left(\left(\left(1-\chi\left(\frac{\epsilon D}{R}\right)\right)f,\phi g\right)+r_\epsilon, \qquad (3.49)$$

with r_ϵ satisfying (3.45).

On the other hand, by the definition of $f^\epsilon(t,x,\xi)$, we have

$$\begin{aligned}
\xi_i f^\epsilon(t,x,\xi)=\frac{\epsilon}{2(2\pi)^d}\int_{\mathbb{R}^d} e^{-iz\xi}\Bigg(&D_{x_i}\psi^\epsilon\left(t,x+\frac{\epsilon z}{2}\right)\overline{\psi^\epsilon\left(t,x-\frac{\epsilon z}{2}\right)}\\
&-\psi^\epsilon\left(t,x+\frac{\epsilon z}{2}\right)D_{x_i}\overline{\psi^\epsilon\left(t,x-\frac{\epsilon z}{2}\right)}\Bigg)\,d\xi; \qquad (3.50)
\end{aligned}$$

hence by (3.44), (3.45), (3.49), and (3.50), we get

(3.51)

$$\begin{aligned}
&\left|\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi(x)\left(1-\chi\left(\frac{\xi}{R}\right)\right)\xi_i f^\epsilon(t,x,\xi)\,dx\,d\xi\right|\\
&=\frac{\epsilon}{2}\left\{\left(W^\epsilon(D_{x_i}\psi^\epsilon,\psi^\epsilon),\phi\left(1-\chi\left(\frac{\xi}{R}\right)\right)\right)-\left(W^\epsilon(\psi^\epsilon,D_{x_i}\psi^\epsilon),\phi\left(1-\chi\left(\frac{\xi}{R}\right)\right)\right)\right\}\\
&\le\left|\left(\phi\epsilon D_{x_i}\psi^\epsilon,\left(1-\chi\left(\frac{\epsilon D}{R}\right)\right)\psi^\epsilon\right)\right|+2r_\epsilon,
\end{aligned}$$

but by (3.34), we have

$$\begin{aligned}\left\|\left(1-\chi\left(\frac{\epsilon D}{R}\right)\right)\psi^\epsilon\right\|_{L^2} &\le \int_{\mathbb{R}^d}\left(1-\chi\left(\frac{\epsilon\xi}{R}\right)\right)^2|\hat{\psi^\epsilon}(t,\xi)|^2\,d\xi\\ &\le \int_{|\xi|\ge\frac{R}{\epsilon}}|\hat{\psi^\epsilon}(t,\xi)|^2\,d\xi\\ &\le \frac{1}{R^2}\int_{|\xi|\ge\frac{R}{\epsilon}}|\epsilon\xi|^2|\hat{\psi^\epsilon}(t,\xi)|^2\,d\xi \le \frac{C}{R^2}.\end{aligned}\tag{3.52}$$

By summing up (3.51) and (3.52), we get

$$\lim_{R\to\infty}\sup_{\epsilon>0}\left|\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi(x)\left(1-\chi\left(\frac{\xi}{R}\right)\right)\xi_i f^\epsilon(t,x,\xi)\,dx\,d\xi\right| = 0,\tag{3.53}$$

while by (3.33) and any $\chi(\xi)$ chosen as that in (3.43), we have trivially that

$$\lim_{\epsilon\to0}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi(x)\chi\left(\frac{\xi}{R}\right)\xi f^\epsilon(t,x,\xi)\,dx\,d\xi = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi(x)\chi\left(\frac{\xi}{R}\right)\xi f(t,dx,d\xi).\tag{3.54}$$

On the other hand, by taking $u=0$ in (3.31), we then infer from (2.2) and the proof of (3.38)–(3.39) that

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|\xi|^2 f(t,dx,d\xi)\le C,$$

which implies that

$$\lim_{R\to\infty}\left|\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi(x)\left(1-\chi\left(\frac{\xi}{R}\right)\right)\xi_i f(t,dx,d\xi)\right| = 0.\tag{3.55}$$

Thus by (3.23) and (3.53), for any fixed $t<T^*$ we pass $R\to\infty$ in (3.54) to obtain

$$\begin{aligned}\lim_{\epsilon\to0}\int_{\mathbb{R}^d}\phi(x)J^\epsilon(t,x)\,dx &= \int_{\mathbb{R}^d}\phi(x)\int_{\mathbb{R}^d}\xi\rho(t,x)\delta(\xi-u(t,x))\,d\xi\,dx\\ &= \int_{\mathbb{R}^d}\phi(x)\rho(t,x)u(t,x)\,dx.\end{aligned}\tag{3.56}$$

This proves (3.41), which completes the proof of the lemma. □

Now we are in a position to complete the proof of Theorem 1.1.

Proof. By Lemma 2.2, (3.23), (3.35), (3.36), and (3.41), to complete the proof of Theorem 1.1, we only need to show part (2) of the theorem. In fact, by (1.10) and Lemma 2.2, for any test function $\phi(x,\xi)\in C_c^\infty(\mathbb{R}^{2d})$, we have

$$\begin{aligned}&\left|\frac{d}{dt}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi(x,\xi)\rho(t,x)\delta(\xi-u(t,x))\,dx\,d\xi\right|\\ &=\left|\frac{d}{dt}\int_{\mathbb{R}^d}\phi(x,u(t,x))\rho(t,x)\,dx\right|\\ &=\left|\int_{\mathbb{R}^d}\{\nabla_\xi\phi(x,u)\partial_t u\rho+\phi(x,u)\partial_t\rho\}\,dx\right|\\ &\le\left(\|\partial_t u\|_{L^2}\|\rho\|_{L^2}+\|\partial_t\rho\|_{L^2}\right)\|\phi\|_{W^{1,\infty}}\\ &\le C\|\phi\|_{H^s},\quad s>d+1\quad\text{for}\quad 0\le t<T^*.\end{aligned}\tag{3.57}$$

This implies that $f(t,x,\xi) \in \mathrm{Lip}([0,T^*), H^{-s}(\mathbb{R}^{2d}))$ for $s > d+1$. On the other hand, to show that $f(t,x,\xi)$ is a distribution solution of (1.6) on $[0,T^*)\times\mathbb{R}^{2d}$, we need only prove that

$$\int_0^{T^*}\int_{\mathbb{R}^d} \{(\partial_t\phi)(t,x,u) + u(\nabla_x\phi)(t,x,u) - E(\nabla_\xi\phi)(t,x,u)\}\,\rho\,dx\,dt = 0 \tag{3.58}$$

for any test function $\phi(t,x,\xi) \in C_c^\infty((0,T^*)\times\mathbb{R}^{2d})$. Notice that by Lemma 2.2, $(\rho(t,x), u(t,x))$ is the unique local smooth solution of (1.10), and we have

$$\int_0^{T^*}\int_{\mathbb{R}^d} \phi(t,x,u)(\partial_t\rho + \mathrm{div}(\rho u))\,dx\,dt = 0 \tag{3.59}$$

and

$$\int_0^{T^*}\int_{\mathbb{R}^d} \nabla_\xi\phi(t,x,u)\{\partial_t u + u\nabla u + E\}\rho\,dx\,dt = 0. \tag{3.60}$$

By summing up (3.59), (3.60) and using integration by parts, we get (3.58). This completes the proof of Theorem 1.1. ∎

Acknowledgments. I am indebted to Professor Brenier, who suggested the idea of using the modulated energy method introduced in [1] for the Vlasov–Poisson system. Later, he mentioned to me the different but related work of Marjolaine Puel [18], who addressed in her Ph.D., with similar methods, the combined classical and neutral limit of the Wigner–Poisson system to the incompressible Euler equations. Also I would like to thank Professors Fanghua Lin, P. A. Markowich, and N. J. Mauser for profitable discussions, with special thanks to Professor Markowich for first introducing me to this problem.

REFERENCES

[1] Y. Brenier, *Convergence of the Vlasov-Poisson system to the incompressible Euler equations*, Comm. Partial Differential Equations, 25 (2000), pp. 737–754.

[2] F. Brezzi and P. A. Markowich, *The three dimensional Wigner-Poisson problem: Existence, uniqueness and approximation*, Math. Methods Appl. Sci., 14 (1991), pp. 35–62.

[3] B. Desjardins, C.-K. Lin, and T.-C. Tso, *Semiclassical limit of the derivative nonlinear Schrödinger equation*, Math. Models Methods Appl. Sci., 10 (2000), pp. 261–285.

[4] I. Gasser and P. A. Markowich, *Quantum hydrodynamics, Wigner transforms and the classical limit*, Asymptot. Anal., 14 (1997), pp. 97–116.

[5] I. Gasser, C. K. Lin, and P. A. Markowich, *A review of dispersive limits of (non)linear Schrödinger-type equations*, Taiwanese J. Math., 4 (2000), pp. 501–529.

[6] P. Gérard, *Microlocal defect measures*, Comm. Partial Differential Equations, 16 (1991), pp. 1761–1794.

[7] P. Gérard, P. A. Markowich, N. J. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 323–379.

[8] J. Ginibre and G. Velo, *On the global Cauchy problem for some nonlinear Schrödinger equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 309–323.

[9] E. Grenier, *Semiclassical limit of the nonlinear Schrödinger equation in small time*, Proc. Amer. Math. Soc., 126 (1998), pp. 523–530.

[10] S. Jin, C. D. Levermore, and D. W. Mclaughlin, *The behavior of solutions of the NLS equation in the semiclassical limit*, in Singular Limits of Dispersive Waves, NATO Adv. Sci. Inst. Ser. B Phys. 320, N. Ercolani, I. Gabitov, C. D. Levermore, and D. Serre, eds., Plenum, New York, 1994, pp. 235–256.

[11] S. Jin, C. D. Levermore, and D. W. Mclaughlin, *The semiclassical limit of the defocusing NLS hierarchy*, Comm. Pure Appl. Math., 52 (1999), pp. 613–654.

[12] L. D. Landau and E. M. Lifschitz, *Lehrbuch der Theoretischen Physik* III. *Quantenmechanik*, Akademie-Verlag, Berlin, 1985.
[13] F. Lin and P. Zhang, *On the hydrodynamic limit of Ginzburg-Landau wave vortices*, Comm. Pure Appl. Math., 55 (2002), pp. 831–856.
[14] P. L. Lions and T. Paul, *Sur les measures de Wigner*, Rev. Mat. Iberoamericana, 9 (1993), pp. 553–618.
[15] M. I. Loffredo and L. M. Morato, *Self-consistent hydrodynamical model for Hell near absolute zero in the frame work of stochastic mechanics*, Phys. Rev. B, 35 (1987), pp. 1742–1747.
[16] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Appl. Math. Sci. 53, Springer-Verlag, New York, 1984.
[17] P. A. Markowich, C. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, New York, 1990.
[18] M. Puel, *Convergence du système de Schrödinger Poisson vers les équations d'Euler*, preprint, 2001.
[19] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Space*, Princeton University Press, Princeton, NJ, 1982.
[20] L. Tartar, *H-measures, a new approach for studying homogenization, oscillations and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 115 (1990), pp. 193–230.
[21] M. Taylor, *Pseudodifferential Operators and Nonlinear PDE*, Birkhäuser Boston, Boston, MA, 1991.
[22] Z. Xin and P. Zhang, *On the global existence of weak solutions to the shallow water equation*, Comm. Pure Appl. Math., 53 (2000), pp. 1411–1433.
[23] E. Wigner, *On the quantum correction for thermodynamic equilibrium*, Phys. Rev., 40 (1932), pp. 749–759.
[24] P. Zhang and Y. Zheng, *Existence and uniqueness of solutions of an asymptotic equation arising from a variational wave equation with general data*, Arch. Ration. Mech. Anal., 155 (2000), pp. 49–83.
[25] P. Zhang, Y. Zheng, and N. J. Mauser, *The limit from Schrödinger-Poisson to Vlasov-Poisson with general data in one dimension*, Comm. Pure Appl. Math., 55 (2002), pp. 582–632.

SIAM J. MATH. ANAL.
Vol. 34, No. 3, pp. 719–735

BOUNDARY DETERMINATION OF CONDUCTIVITIES AND RIEMANNIAN METRICS VIA LOCAL DIRICHLET-TO-NEUMANN OPERATOR*

HYEONBAE KANG† AND KIHYUN YUN†

Abstract. We consider the inverse problem to identify an anisotropic conductivity from the Dirichlet-to-Neumann (DtN) map. We first find an explicit reconstruction of the boundary value of less regular anisotropic (transversally isotropic) conductivities and their derivatives. Based on the reconstruction formula, we prove Hölder stability, up to isometry, of the inverse problem using a local DtN map.

Key words. inverse boundary value problem, Dirichlet-to-Neumann map, anisotropic conductivity, boundary determination

AMS subject classification. 35R30

PII. S0036141001395042

1. Introduction and statements of results.

The results of this paper are twofold. We first find an explicit reconstruction of boundary values of anisotropic conductivities and their derivatives. We then derive Hölder stability estimates for the inverse problem to identify Riemannian metrics (up to isometry) on the boundary via the local Dirichlet-to-Neumann (DtN) map using the same ideas and methods.

Boundary reconstruction. Let Ω be a bounded domain in $\mathbb{R}^n$ $(n \geq 2)$ with the smooth boundary. If we recover the conductivity up to mth derivatives, then it is enough to assume that $\partial\Omega$ is C^{m+2}-smooth. We consider the inverse problem of identifying the positive definite symmetric matrix $\gamma = (\gamma^{ij})$ entering the equation

$$L_\gamma u := \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(\gamma^{ij}\frac{\partial u}{\partial x_j}\right) = 0 \quad \text{in } \Omega \tag{1.1}$$

by the DtN map. The DtN map $\Lambda_\gamma : H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega)$ is defined to be

$$\langle \Lambda_\gamma f, h\rangle = \int_\Omega (\gamma\nabla u)\cdot\nabla v dx, \quad f, h \in H^{1/2}(\partial\Omega),$$

where $u \in H^1(\Omega)$ is the solution to (1.1) with the Dirichlet data $u|_{\partial\Omega} = f$, and $v \in H^1(\Omega)$ is such that $v|_{\partial\Omega} = h$. Here $\langle\ ,\ \rangle$ denotes the $H^{-1/2}(\partial\Omega)$-$H^{1/2}(\partial\Omega)$ pairing.

In this paper we are first concerned with an explicit reconstruction of the conductivity at the boundary. Quite recently, Nakamura and Tanuma [13, 14] obtained an explicit formula to reconstruct conductivity and its normal derivatives at the boundary. We will discuss more about their formula since the first main result of this paper is an improvement of it.

*Received by the editors September 10, 2001; accepted for publication (in revised form) July 4, 2002; published electronically January 7, 2003. This work was supported in part by KOSEF 98-0701-03-5 and 2000-1-10300-001-1.

http://www.siam.org/journals/sima/34-3/39504.html

†School of Mathematical Sciences, Seoul National University, Seoul 151-747, Korea (hkang@math.snu.ac.kr, khyun@math.snu.ac.kr).

Let us suppose that $\partial\Omega$ is flat around $x_0 = 0 \in \partial\Omega$, namely, there exists $\delta > 0$ such that

$$\Omega \cap B_\delta(0) = \{x = (x', x_n) \in B_\delta(0) \mid x_n > 0\}, \tag{1.2}$$

and in $\overline{\Omega} \cap B_\delta(0)$, γ is given by

$$\gamma = \begin{pmatrix} & & & 0 \\ & \gamma^{ij} & & \vdots \\ & & & 0 \\ 0 & \cdots & 0 & \gamma^{nn} \end{pmatrix}. \tag{1.3}$$

In fact, using the boundary normal coordinates [9], we can locally transform the general conductivity γ to one of the form (1.2). One can even take $\gamma^{nn} = 1$. Here we keep γ^{nn} in order to see how much we can recover.

Let $t' \in \mathbb{R}^{n-1}$, i.e., $(t', 0)$ is a tangent vector to $\partial\Omega$ at x_0. Let $\eta(x') \in C_0^\infty(\mathbb{R}^{n-1})$ be such that

$$0 \le \eta \le 1, \quad \|\eta\|_{L^2} = 1, \quad \text{supp } \eta \subset \{|x'| < 1\}.$$

For each large positive integer N, let

$$\phi_N(x') = \exp(iNx' \cdot t')\eta(N^{1/2}x'), \tag{1.4}$$

and for $z \in \partial\Omega \cap B_\delta(0)$, let

$$\phi_N^z(x') = \phi_N(x' - z'). \tag{1.5}$$

The function ϕ_N plays the role of Dirichlet data and test functions. Observe that ϕ_N oscillates rapidly as N becomes large. Kohn and Vogelius first used rapidly oscillating boundary data in their proof of uniqueness of the boundary determination [6]. The use of explicit functions such as ϕ_N for boundary reconstruction is due to Brown [3] and Nakamura and Tanuma [13].

Let $\gamma^k \in C^m(\overline{\Omega})$ be an anisotropic conductivity such that

$$\gamma^k(x) := \gamma(x', 0) + \partial_n\gamma(x', 0)x_n + \cdots + \frac{1}{(k-1)!}\partial_n^{k-1}\gamma(x', 0)x_n^{k-1} \tag{1.6}$$

near $\partial\Omega \cap B_\delta(0)$. Let

$$C_\gamma(z) := \sqrt{\gamma^{nn}(z)^{-1}\sum_{i,j=1}^{n-1}\gamma^{ij}(z)t_it_j}. \tag{1.7}$$

Nakamura and Tanuma proved that for $k \le \frac{m}{2}$,

$$\begin{aligned} &\lim_{N\to\infty} N^{\frac{n-3}{2}+k}\langle(\Lambda_\gamma - \Lambda_{\gamma^k})\phi_N^z, \overline{\phi_N^z}\rangle \\ &\quad = C_k C_\gamma(z)^{-a_n-1}\left(\sum_{i,j=1}^{n-1}\partial_{x_n}^k\gamma^{ij}(z)t_it_j + C_\gamma(z)^2\partial_{x_n}^k\gamma^{nn}(z)\right) \end{aligned} \tag{1.8}$$

for some explicit constant C_k. (Even if they wrote the formula only for the isotropic γ, the proof gives (1.8).) If γ is isotropic and $|t'| = 1$, then $C_\gamma \equiv 1$, and hence the formula (1.8) reads

$$\lim_{N\to\infty} N^{\frac{n-3}{2}+k}\langle(\Lambda_\gamma - \Lambda_{\gamma^k})\phi_N^z, \overline{\phi_N^z}\rangle = 2C_k\partial_{x_n}^k\gamma(z). \tag{1.9}$$

However, the reconstruction formula (1.8) is valid only for $k \le \frac{m}{2}$. Moreover, in the inductive reconstruction (1.8) or (1.9), it is required to know $\partial_{x_n}^{k-1}\gamma(x)$ for all $x \in \partial\Omega \cap B_\delta(0)$ in order to recover $\partial_{x_n}^k\gamma(z)$. It is also worth noting that in the formula (1.8) or (1.9), stable recovery of the tangential derivatives does not seem possible: If we take $N^{\frac{n-3}{2}+k}\langle(\Lambda_\gamma - \Lambda_{\gamma_k})\phi_N^z, \overline{\phi_N^z}\rangle$ as an approximation of $2C_k\partial_{x_n}^k\gamma(z)$ in (1.9), then its tangential derivatives do not seem to be good approximations of those of $2C_k\partial_{x_n}^k\gamma(z)$. It is our intention to improve these points.

The reason for the above-mentioned drawbacks in the formula (1.8) is that $N^{1/2}$ is used in the definition (1.4). We use instead the following boundary data:

$$\phi_N(x') = \exp(iNx'\cdot t')\eta(N^{\alpha_1}x_1, \dots, N^{\alpha_{n-1}}x_{n-1}), \tag{1.10}$$

where α_j's are specified shortly. The definition (1.10) amounts to assigning each partial differential operator $\frac{\partial}{\partial x_j}$ with the weight α_j $(j = 1, \dots, n)$ so that we can distinguish each direction x_j. The numbers α_j are chosen as follows: throughout this paper conductivities under consideration are $C^{m,p}$-smooth $(m \ge 0,\ p > 0)$, and $C^{m,p}$, m nonnegative integer and $0 \le p \le 1$, denotes the usual Hölder space. Choose λ so that $\lambda = \frac{1}{l}$ for some integer l and satisfies the following: if $m \ge 1$, then

$$\lambda < p, \quad (1 - m^{n-1}\lambda)(m + p) \ge m + \lambda. \tag{1.11}$$

We then define a multi-index α by

$$\alpha = (\alpha_1, \dots, \alpha_n) := (1 - m^{n-1}\lambda, 1 - m^{n-2}\lambda, \dots, 1 - m\lambda, 1). \tag{1.12}$$

If $m = 0$, choose λ so that $\lambda < p$ and define α by

$$\alpha = (\alpha_1, \dots, \alpha_n) := (1 - (n-1)\lambda, 1 - (n-2)\lambda, \dots, 1). \tag{1.13}$$

We choose α and λ in this way so that they possess the following properties: $|\alpha_i - \alpha_j| \ge \lambda$ if $i \ne j$. If a and b are multi-indices with $|a| \le m$ and $|b| \le m$, then $a\cdot\alpha \ne b\cdot\alpha$ if and only if $a \ne b$. Thanks to these properties, we can define a linear ordering of multi-indices: Let $a = (a_1, \dots, a_n)$ and $b = (b_1, \dots, b_n)$ be two multi-indices. We define $a < b$ if $a\cdot\alpha < b\cdot\alpha$. Using the linear ordering, we are able to recover γ and its derivatives inductively.

For the function η in the definition (1.10), we further assume that for each $a' = (a_1, \dots, a_{n-1})$ with $|a'| \le m$

$$\int_{|y'|\le 1} (y')^{a'}\eta(y')^2 dy' \ne 0,$$

and we define $C(a)$ for a multi-index $a = (a', a_n)$ to be

$$C(a) := \frac{1}{a!}\int_0^\infty y_n^{a_n} e^{-2y_n} dy_n \int_{|y'|\le 1} (y')^{a'}\eta(y')^2 dy'. \tag{1.14}$$

For a given anisotropic conductivity γ and a multi-index a with $|a| \le m$, define $\gamma^{a,z}$ to be a positive definite matrix-valued smooth function on Ω such that

$$\gamma^{a,z}(x) := \sum_{b<a} \frac{\partial^b \gamma(z)}{b!}(x-z)^b \quad \text{near } z. \tag{1.15}$$

Then the DtN map $\Lambda_{\gamma^{a,z}}$ corresponding to $\gamma^{a,z}$ is well defined. If $a = 0$, let $\Lambda_{\gamma^0} = 0$. Here and throughout this paper $\partial^b\gamma$ denotes $\partial^b\gamma = \partial_{x_1}^{b_1} \cdots \partial_{x_n}^{b_n}\gamma$.

Then we have the following reconstruction formula.

THEOREM 1.1. *Suppose that $\gamma \in C^{m,p}(\overline{\Omega} \cap B_\delta(0))$. For $z = (z', 0) \in \partial\Omega \cap B_\delta(0)$, a multi-index $a = (a', a_n)$, and $k \le m$, we have*

$$\begin{aligned} &N^{-2+|\alpha|+a\cdot\alpha}\langle(\Lambda_\gamma - \Lambda_{\gamma^{a,z}})\phi_N^z, \overline{\phi_N^z}\rangle \\ &\quad = C(a)C_\gamma(z)^{-a_n-1}\left(\sum_{i,j=1}^{n-1} \partial^a\gamma^{ij}(z)t_it_j + C_\gamma(z)^2\partial^a\gamma^{nn}(z)\right) + O(N^{-\lambda}), \end{aligned} \tag{1.16}$$

where $O(N^{-\lambda})$ is independent of z. If $a = 0$, then $C(0) = \frac{1}{2}$, and hence we have

$$N^{-2+|\alpha|}\langle\Lambda_\gamma\phi_N^z, \overline{\phi_N^z}\rangle = \sqrt{\gamma^{nn}(z)\sum_{i,j=1}^{n-1}\gamma^{ij}(z)t_it_j} + O(N^{-\lambda}). \tag{1.17}$$

The formula (1.16) says that the boundary values of γ and its derivatives up to order m can be recovered in a stable way (modulo γ^{nn}-terms).

In particular, if γ is isotropic, namely, $\gamma = \gamma(\delta_{ij})$, and $|t'| = 1$, then $C_z \equiv 1$, and hence we have the following corollary.

COROLLARY 1.2. *If $\gamma \in C^{m,p}(\overline{\Omega} \cap B_\delta(0))$ is an isotropic conductivity, then for all multi-index a with $|a| \le m$, we have*

$$N^{-2+|\alpha|+a\cdot\alpha}\langle(\Lambda_\gamma - \Lambda_{\gamma^{a,z}})\phi_N^z, \overline{\phi_N^z}\rangle = C(a)\partial^a\gamma(z) + O(N^{-\lambda}). \tag{1.18}$$

We note that Brown proved a reconstruction formula for γ and the normal derivative on $\partial\Omega$ [3]. In [2], Alessandrini and Gaburro considered reconstruction of special types of anisotropic conductivity.

It turns out that a slight variance of the reconstruction (1.18) gives an interesting stability result, which is the second subject of this paper.

Boundary determination of Riemannian metrics-stability. Let Ω be a bounded domain in $\mathbb{R}^n$ $(n \ge 3)$ with the smooth boundary. We consider the inverse problem of identifying a Riemannian metric or an anisotropic conductivity at the boundary $\partial\Omega$ via the (local) DtN map. Let (g_{ij}) be a Riemannian metric on $\overline{\Omega}$ and $g = (g^{ij}) := (g_{ij})^{-1}$. Then the corresponding DtN map $\Lambda_g : H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega)$ is defined to be

$$\langle\Lambda_g f, h\rangle = \int_\Omega (|g|^{-1/2}g\nabla u)\cdot\nabla v dx, \quad f, h \in H^{1/2}(\partial\Omega), \tag{1.19}$$

where $u \in H^1(\Omega)$ is the solution to the problem

$$\begin{aligned} \Delta_g u := |g|^{1/2}\sum_{i,j=1}^n \frac{\partial}{\partial x_i}\left(|g|^{-1/2}g^{ij}\frac{\partial u}{\partial x_j}\right) = 0 \quad &\text{in } \Omega, \\ u = f \quad &\text{on } \partial\Omega, \end{aligned}$$

and $v \in H^1(\Omega)$ is such that $v|_{\partial\Omega} = h$. Here $|g|$ denotes the determinant of g.

There is a well-known obstacle in identifying g: Let $\Psi : \overline{\Omega} \to \overline{\Omega}$ be a C^1 diffeomorphism which is the identity on $\partial\Omega$. Then it is well known that

$$\Lambda_{\Psi^* g} = \Lambda_g,$$

where $\Psi^* g$ is the pull-back of g. So the general conjecture of the uniqueness in three dimensions is that if $\Lambda_{g_1} = \Lambda_{g_2}$, then there exists a diffeomorphism Ψ on $\overline{\Omega}$ such that $\Psi|_{\partial\Omega}$ is the identity on $\partial\Omega$ and

$$\Psi^* g_2 = g_1.$$

Lee and Uhlmann proved the conjecture for three dimensions under some restrictions when conductivities are real analytic in $\overline{\Omega}$ [9]. Recently, Lassas and Uhlmann extended the result to the case when g is real analytic up to a portion of $\partial\Omega$ and removed the restrictions [8]. There are also similar kinds of uniqueness theorems for two dimensions. See [15, 10, 8]. When γ is a scalar function, i.e., γ is isotropic, the inverse problem has been extensively studied [4, 5, 6, 7, 11, 12, 16, 17].

In this paper we prove the following Hölder stability estimates for the boundary determination: Let Γ be an open connected subset of $\partial\Omega$. Define the localized DtN map Λ_g^Γ by

$$\Lambda_g^\Gamma(f) := \Lambda_g(f)|_\Gamma, \quad f \in H^{1/2}(\partial\Omega), \quad \operatorname{supp}(f) \subset \Gamma.$$

We will use the following notation: Let f be a C^k function in a neighborhood of a compact set K. Then $\|f\|_{C_E^k(K)} := \sum_{|\alpha|=0}^k \sup_{x\in K} |\partial^\alpha f(x)|$. So, $C_E^k(\partial\Omega)$-norm involves not only the tangential derivatives but also the normal derivative.

THEOREM 1.3. *Suppose that g_1 and g_2 are Riemannian metrics on a domain Ω such that they are $C^{m,p}$ $(m \geq 1, p > 0)$ in a neighborhood of Γ and the $C^{m,p}$-norms are bounded by M and*

$$g_j \xi \cdot \xi \geq A|\xi|^2 \quad \text{for all } \xi \in \mathbb{R}^n \quad (j = 1, 2). \tag{1.20}$$

Let K be a compact subset of Γ. Then there are a neighborhood U of Γ, a C^m diffeomorphism Ψ in $U \cap \overline{\Omega}$ with $\Psi|_\Gamma$ = Identity, and a positive constant $C = C(m, p, \Gamma, K, A, M)$ such that for $k = 0, 1, \ldots, m$,

$$\|g_2 - \Psi^* g_1\|_{C_E^k(K)} \leq C \|\Lambda_{g_1}^\Gamma - \Lambda_{g_2}^\Gamma\|^{2^{-k/\lambda}}. \tag{1.21}$$

The norm on the right-hand side of (1.21) *is the operator norm from $H^{1/2}(\Gamma)$ into $H^{-1/2}(\Gamma)$.*

Thus the Riemannian metrics can be recovered at the boundary in a stable way via the local DtN map.

Using the boundary normal coordinates, we may assume that $\partial\Omega$ is flat around $x_0 = 0 \in \partial\Omega$, and g is given by

$$g = \begin{pmatrix} & & & 0 \\ & g^{ij} & & \vdots \\ & & & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{1.22}$$

We prove stability estimates for the Riemannian metrics of the form (1.22) by using methods similar to Theorem 1.1. Then Theorem 1.3 follows.

As was observed in [9], if we take

$$g_{ij} := |\gamma|^{1/(n-2)}\gamma^{-1} \quad (n \geq 3),$$

then $\gamma = |g|^{-1/2}g$ ($g = (g_{ij})^{-1}$) and $\Lambda_g = \Lambda_\gamma$. Hence, we have the same stability estimates for anisotropic conductivities.

A similar proof yields the following stability for the boundary determination of isotropic conductivities.

THEOREM 1.4. *Suppose that γ_1 and γ_2 are isotropic conductivities. Let m, p, Γ, K, A, M ($m \geq 0$) be as before. Then there exists a positive constant $C = C(m, \Gamma, K, A, M)$ such that for $k = 0, 1, \dots, m$,*

$$\|\gamma_2 - \gamma_1\|_{C_E^k(K)} \leq C\|\Lambda_{\gamma_1}^\Gamma - \Lambda_{\gamma_2}^\Gamma\|^{2^{-k/\lambda}}. \tag{1.23}$$

Alessandrini proved the following stability for isotropic conductivities using singular solutions [1] (see also [18]):

$$\|\gamma_2 - \gamma_1\|_{C_E^k(\partial\Omega)} \leq C\|\Lambda_{\gamma_1} - \Lambda_{\gamma_2}\|^{\frac{1}{2k+1}}.$$

This stability estimate is better than (1.23). However, the stability estimate (1.23) uses the *local* DtN map.

This paper is organized as follows: In section 2, we construct approximate solutions of $L_\gamma u = 0$ with $u|_{\partial\Omega} = \phi_N^z$ on which the proof of Theorem 1.1 is based. The proof of Theorem 1.1 is given in section 3. Theorem 1.3 is proved in section 4.

2. Approximate solutions. Suppose that Ω and γ are of the forms (1.2) and (1.3) and that $\gamma \in C^{m,p}(\overline{\Omega} \cap B_\delta(0))$ for some integer $m \geq 0$. Let

$$\Omega_N := \left\{ x \mid |x_j| < N^{-\alpha_j} \ (j = 1, \dots, n-1),\ 0 \leq x_n < \frac{1}{\sqrt{N}} \right\}.$$

The following lemma and its proof are based on an idea in [13].

LEMMA 2.1. *For each integer N there is an approximate solution Φ_N of the form*

$$\Phi_N(x) = \exp(iNx' \cdot t') \exp(-C_\gamma(z)Nx_n) \sum_{k=0}^{m/\lambda} N^{-k\lambda} v_k(y) \tag{2.1}$$

$$(y_j = N^{\alpha_j} x_j, \quad j = 1, \dots, n), \tag{2.2}$$

where $v_0(y', y_n) = \eta(y')$, and $v_l(y', y_n)$ are polynomials of y_n whose coefficients are C^∞ functions of y' compactly supported in $\{|y'| < 1\}$, so that Φ_N satisfies

$$\Phi_N|_{\partial\Omega} = \phi_N^z$$

and

$$|\nabla \cdot (\gamma \nabla \Phi_N)(x)| \leq CN^{(2-m)-\lambda} p(y_n) e^{-C_\gamma(z)y_n} \quad \text{for all } x \in \Omega_N \tag{2.3}$$

for some constant $C = C(m)$. Here $p(y_n)$ is a polynomial with positive coefficients.

Proof. Without loss of generality, assume that $z = 0$. Put $C_0 := C_\gamma(0)$. We seek a solution $\Phi_N(x)$ of the form

$$\Phi_N(x) = \exp(iNx' \cdot t') V(N^{\alpha_1} x_1, \dots, N^{\alpha_n} x_n).$$

Then straightforward computations show that

$$\begin{aligned}
\nabla_x(\gamma(\nabla_x\Phi_N)) &= \sum_{i,j=1}^{n} \partial_{x_i}(\gamma^{ij}\partial_{x_j}\Phi_N) \\
&= \left[\sum_{i,j=1}^{n-1} \gamma^{ij}\partial_{x_i}\partial_{x_j} + \gamma^{nn}\partial_{x_n}^2 + \sum_{i,j=1}^{n-1} \partial_{x_i}\gamma^{ij}\partial_{x_j} + \partial_{x_n}\gamma^{nn}\partial_{x_n}\right]\Phi_N \\
&= \exp(iNx'\cdot t')\left[-N^2\sum_{i,j=1}^{n-1}\gamma^{ij}t_it_j + \sqrt{-1}N\sum_{i,j=1}^{n-1}\gamma^{ij}(t_i\partial_{x_j} + t_j\partial_{x_i})\right. \\
&\quad + \sum_{i,j=1}^{n-1}\gamma^{ij}\partial_{x_i}\partial_{x_j} + \gamma^{nn}\partial_{x_n}^2 + \sqrt{-1}N\sum_{i,j=1}^{n-1}(\partial_{x_i}\gamma^{ij})t_j \\
&\quad \left. + \sum_{i,j=1}^{n-1}(\partial_{x_i}\gamma^{ij})\partial_{x_j} + (\partial_{x_n}\gamma^{nn})\partial_{x_n}\right] V(N^{\alpha_1}x_1,\dots,N^{\alpha_n}x_n).
\end{aligned}$$

After the scaling (2.2), we have $\partial_{x_j} = N^{\alpha_j}\partial_{y_j}$ $(i = 1, 2, \dots, n)$, and hence

$$\begin{aligned}
\nabla_x(\gamma(\nabla_x\Phi_N)) = \exp(iNx'\cdot t')\Bigg[&N^2\left(\gamma^{nn}\partial_{y_n}^2 - \sum_{i,j=1}^{n-1}\gamma^{ij}t_it_j\right) \\
&+ 2\sum_{j=1}^{n-1} N^{1+\alpha_j}\left(\sum_{i=1}^{n-1}\gamma^{ij}t_i\right)\partial_{y_j} \\
&+ \sum_{i,j=1}^{n-1} N^{\alpha_i+\alpha_j}\gamma^{ij}\partial_{y_i}\partial_{y_j} \\
&+ N\left(\sqrt{-1}\sum_{i,j=1}^{n-1}(\partial_{x_i}\gamma^{ij})t_j + (\partial_{x_n}\gamma^{nn})\partial_{y_n}\right) \\
&+ \sum_{j=1}^{n-1} N^{\alpha_j}\left(\sum_{i=1}^{n-1}\partial_{x_i}\gamma^{ij}\right)\partial_{y_j}\Bigg] V(y', y_n).
\end{aligned} \tag{2.4}$$

Note that all the powers of N in the formula (2.4) are of the form $2 - k\lambda$ for some integer k with $0 \le k \le 2/\lambda$.

We now expand γ in Taylor series in Ω_N:

$$\gamma(x) = \sum_{|a|\le m}\frac{1}{a!}\partial^a\gamma(0)x^a + O(|x|^{m+p}).$$

By the condition (1.11) imposed on λ, we have

$$\alpha_1(m+p) \ge m + \lambda. \tag{2.5}$$

Thus, after the scaling (2.2), we have

$$\gamma(x) = \sum_{|a|\le m}\frac{1}{a!}\partial^a\gamma(0)N^{-\alpha\cdot a}y^a + E_1(y), \tag{2.6}$$

where

$$|E_1(y)| \le CN^{-\alpha_1(m+p)} \sum_{k=0}^{m+1} y_n^k \le CN^{-m-\lambda} \sum_{k=0}^{m+1} y_n^k. \tag{2.7}$$

Similarly, we have, for $j = 1, 2, \dots, n$,

$$\partial_{x_j}\gamma(x) = \sum_{|a|\le m-1} \frac{1}{a!}\partial^a \partial_{x_j}\gamma(0) N^{-\alpha\cdot a} y^a + E_2(y), \tag{2.8}$$

where

$$|E_2(y)| \le CN^{-m+1-\lambda} \sum_{k=0}^{m} y_n^k. \tag{2.9}$$

Note that this expansion holds uniformly for all $x \in \Omega_N$ and hence for all $y \in \{|y'| < 1,\ 0 \le y_n \le N^{\frac{1}{2}}\}$. Note also that the powers of N in the expansions in (2.6) and (2.8) are of the form $-k\lambda$ for some integer k.

It then follows from (2.4), (2.6), and (2.8) that

$$\nabla_x(\gamma(\nabla_x \Phi_N)) = \exp(iNx'\cdot t') \left[\sum_{k=0}^{m/\lambda} N^{2-k\lambda} L_k + L_R\right] V(y', y_n), \tag{2.10}$$

where L_k are at most second order differential operators in y' and y_n whose coefficients are polynomials in y' and y_n, and L_R is also a second order differential operator in y' and y_n whose coefficients are of the form $O(N^{2-m-\lambda})\times$polynomial in y_n with positive coefficients, and

$$L_0 = \gamma^{nn}(0)\partial_{y_n}^2 - \sum_{i,j=1}^{n-1} \gamma^{ij}(0) t_i t_j. \tag{2.11}$$

We look for $V(y', y_n)$ of the form

$$V(y', y_n) = \sum_{k=0}^{m/\lambda} N^{-k\lambda} V_k.$$

We have from (2.10) that

$$\begin{aligned} \nabla_x(\gamma(\nabla_x\Phi_N)) &= \exp(iNx'\cdot t') \left[\left(\sum_{k=0}^{m/\lambda} N^{2-k\lambda} L_k\right)\left(\sum_{j=0}^{m/\lambda} N^{-j\lambda} V_j\right) + L_R V\right] \\ &= \exp(iNx'\cdot t') \left[\sum_{l=0}^{m/\lambda} N^{2-l\lambda} \sum_{k+j=l} L_k V_j + E\right], \end{aligned} \tag{2.12}$$

where

$$E := \sum_{l=m/\lambda+1}^{2m/\lambda} N^{2-l\lambda} \sum_{k+j=l} L_k V_j + L_R V. \tag{2.13}$$

We solve the system of differential equations

$$\sum_{k+j=l} L_k V_j = 0 \quad (l = 0, 1, 2, \dots, m/\lambda), \tag{2.14}$$

namely,

$$\begin{aligned} L_0 V_0 &= 0, \\ L_0 V_1 + L_1 V_0 &= 0, \\ &\cdots \\ L_0 V_{m/\lambda} + \cdots + L_{m/\lambda} V_0 &= 0, \end{aligned}$$

with the boundary conditions

$$\begin{aligned} V_0|_{y_n=0} &= \eta_N(x') = \eta(y'), \\ V_l|_{y_n=0} &= 0 \quad (l = 1, \dots, m/\lambda). \end{aligned}$$

We remark that this boundary value problem is underdetermined. Because of (2.11), this system of equations can be solved iteratively from top to bottom:

$$\begin{aligned} V_0(y', y_n) &= \eta(y') \exp(-C_0 y_n), \\ V_1(y', y_n) &= \sum_{k=0}^{1} P_1^k(y') y_n^k \exp(-C_0 y_n), \\ &\cdots \\ V_{m/\lambda}(y', y_n) &= \sum_{k=1}^{m/\lambda} P_{m/\lambda}^k(y') y_n^k \exp(-C_0 y_n) \end{aligned}$$

for some N_j $(j = 1, \dots, m/\lambda)$, where $P_j^k(y')$ are C^∞ functions supported in $\{|y'| < 1\}$. It then follows from (2.12) that

$$\nabla_x(\gamma(\nabla_x \Phi_N)) = \exp(iNx' \cdot t')E.$$

Recall that the coefficients of L_R are of the form $O(N^{2-m-\lambda}) \times$ polynomial in y_n. Thus there exists $C = C(m)$ such that

$$|E| \le C N^{2-m-\lambda} p(y_n) \exp(-C_0 y_n)$$

for some polynomial p. This completes the proof. □

The following lemma can be proved by straightforward computations. Recall that ϕ_N is defined in (1.10).

LEMMA 2.2. *For each $s \ge 0$, there exists a constant C_s such that*

$$\|\phi_N\|_{H^s(\partial\Omega)} + \|\Phi_N\|_{H^{s+1/2}(\Omega_N)} \le C_s N^{s+\frac{1}{2}-\frac{|\alpha|}{2}}. \tag{2.15}$$

For each multi-index a, there exists a constant C_a such that

$$\|x^a \nabla \Phi_N\|_{L^2(\Omega_N)} \le C_a N^{1-a\cdot\alpha-\frac{|\alpha|}{2}}. \tag{2.16}$$

3. Proof of Theorem 1.1. In this section we prove Theorem 1.1. Our proof is parallel to that of [13].

Without loss of generality we assume that $z = 0$. Put $C_0 := C_\gamma(0)$ for convenience. Let $\zeta(x_n) \in C^\infty([0,\infty))$ be such that $\zeta(x_n) = 1$ for $0 \le x_n \le 1/2$ and 0 for $1 \le x_n$. Put

$$\zeta_N(x_n) = \zeta(\sqrt{N}x_n).$$

Let a be a multi-index such that $|a| \le m$. Let $u_N \in H^1(\Omega)$ be the solution of

$$\begin{aligned} \nabla_x(\gamma\nabla_x u_N) &= 0 \quad \text{in } \Omega, \\ u_N|_{\partial\Omega} &= \phi_N, \end{aligned}$$

and let $v_N \in H^1(\Omega)$ be the solution of

$$\begin{aligned} \nabla_x(\gamma^a\nabla_x v_N) &= 0 \quad \text{in } \Omega, \\ v_N|_{\partial\Omega} &= \phi_N. \end{aligned}$$

Here $\gamma^a = \gamma^{a,0}$. Let Φ_N and Ψ_N be the extensions of ϕ_N given in Lemma 2.1 corresponding to γ and γ^a, respectively. Note that since $\langle \Lambda_{\gamma^a}\phi_N, \overline{\phi}_N\rangle$ is real, we have

$$\langle \Lambda_{\gamma^a}\phi_N, \overline{\phi_N}\rangle = \langle \overline{\Lambda_{\gamma^a}\phi_N}, \phi_N\rangle,$$

and hence

$$\begin{aligned} &\langle (\Lambda_\gamma - \Lambda_{\gamma^a})\phi_N, \overline{\phi_N}\rangle \\ &= \int_\Omega (\gamma\nabla_x u_N)\cdot\nabla_x\overline{(\zeta_N\Psi_N)}dx - \int_\Omega (\gamma^a\nabla_x\overline{v_N})\cdot\nabla_x(\zeta_N\Phi_N)dx. \end{aligned}$$

Put

$$u_N := \Phi_N + s_N \quad \text{and} \quad v_N := \Psi_N + r_N. \tag{3.1}$$

Then we have

$$\begin{aligned} &\langle (\Lambda_\gamma - \Lambda_{\gamma^a})\phi_N, \overline{\phi_N}\rangle \\ &= \int_\Omega \left[(\gamma\nabla_x\Phi_N)\cdot\nabla_x\overline{(\zeta_N\Psi_N)} - (\gamma^a\nabla_x\overline{\Psi_N})\cdot\nabla_x(\zeta_N\Phi_N)\right] dx \\ &\quad + \int_\Omega (\gamma\nabla_x s_N)\cdot\nabla_x\overline{(\zeta_N\Psi_N)}dx - \int_\Omega (\gamma^a\nabla_x\overline{r_N})\cdot\nabla_x(\zeta_N\Phi_N)dx \\ &:= I + II + III. \end{aligned} \tag{3.2}$$

We estimate I, II, and III separately.

Estimates of I. Put

$$\Omega'_N := \left\{ x : \ |x'| \le N^{-\alpha_j} \ (j = 1,\dots,n-1), \ \frac{1}{2\sqrt{N}} \le x_n \le \frac{1}{\sqrt{N}} \right\}.$$

Since $\zeta_N = 1$ on $0 \le x_n \le \frac{1}{2\sqrt{N}}$, we can rewrite I as

$$\begin{aligned} I &= \int_{\Omega_N\setminus\Omega'_N} (\gamma - \gamma^a)\nabla_x\Phi_N\cdot\nabla_x\overline{\Psi_N}dx \\ &\quad + \int_{\Omega'_N} \left[(\gamma\nabla_x\Phi_N)\cdot\nabla_x\overline{(\zeta_N\Psi_N)} - (\gamma^a\nabla_x\overline{\Psi_N})\cdot\nabla_x(\zeta_N\Phi_N)\right] dx \\ &:= I_1 + I_2. \end{aligned}$$

By (2.1), there exists a constant C such that

$$|\Psi_N| + |\Phi_N| \le C \exp(-C_0 N x_n).$$

Therefore it is easy to see that

$$|I_2| \le C \exp\left(-\frac{C_0}{2} N^{\frac{1}{2}}\right). \tag{3.3}$$

From (2.1), we get

$$\nabla_x \Phi_N = \left[N \begin{pmatrix} it' \\ -C_0 \end{pmatrix} \exp(iNx' \cdot t') \eta_N(x') + O(N^{1-\lambda}) \right] \exp(-C_0 N x_n).$$

Likewise, we have

$$\nabla_x \Psi_N = \left[N \begin{pmatrix} it' \\ -C_0 \end{pmatrix} \exp(iNx' \cdot t') \eta_N(x') + O(N^{1-\lambda}) \right] \exp(-C_0 N x_n).$$

It then follows that

$$\begin{aligned} I_1 = N^2 \int_0^{\frac{1}{2\sqrt{N}}} \int_{|x'| \le \frac{1}{\sqrt{N}}} & \left((\gamma(x) - \gamma^a(x)) \begin{pmatrix} it' \\ -C_0 \end{pmatrix} \right) \cdot \begin{pmatrix} -it' \\ -C_0 \end{pmatrix} \\ & \times e^{-2C_0 N x_n} \eta_N(x')^2 dx' dx_n \\ & + O(N^{2-\lambda}) \int_0^{\frac{1}{2\sqrt{N}}} \int_{|x'| \le \frac{1}{\sqrt{N}}} e^{-2C_0 N x_n} |\gamma(x) - \gamma^a(x)| dx' dx_n. \end{aligned}$$

Note that if two multi-indices a and b satisfy $a < b$, then

$$a \cdot \alpha + \lambda \le b \cdot \alpha.$$

Thus, applying the change of variables $y_j = N^{\alpha_j} x_j$, $j = 1, \dots, n$, we obtain

$$\begin{aligned} \gamma(x) - \gamma^a(x) &= \frac{\partial^a \gamma(0)}{a!} x^a + \sum_{b > a, |b| \le m} \frac{\partial^b \gamma(0)}{b!} x^b + O(|x|^{m+1}) \\ &= \frac{\partial^a \gamma(0)}{a!} N^{-a \cdot \alpha} y^a + O(N^{-a \cdot \alpha - \lambda}) \sum_{k=0}^{m+1} y_n^k. \end{aligned}$$

Therefore, we have

$$\begin{aligned} I_1 = {} & \frac{N^{2-a\cdot\alpha-|\alpha|}}{a!} \int_0^{\frac{\sqrt{N}}{2}} \int_{|y'| \le 1} y^a \partial^a \gamma(0) \begin{pmatrix} it' \\ -C_0 \end{pmatrix} \cdot \begin{pmatrix} -it' \\ -C_0 \end{pmatrix} e^{-2C_0 y_n} \eta(y')^2 dy' dy_n \\ & + O(N^{2-a\cdot\alpha-\lambda-|\alpha|}) \int_0^{\frac{\sqrt{N}}{2}} \int_{|y'| \le 1} y^a e^{-2C_0 y_n} [\eta(y')^2 + 1] \sum_{k=0}^{m+1} y_n^k dy' dy_n \\ = {} & N^{2-a\cdot\alpha-|\alpha|} \left(\sum_{i,j=1}^{n-1} \partial^a \gamma^{ij}(0) t_i t_j + C_0^2 \partial^a \gamma^{nn}(0) \right) \\ & \times \frac{1}{a!} \int_0^{\infty} \int_{|y'| \le 1} y^a e^{-2C_0 y_n} \eta(y')^2 dy' dy_n + O(N^{2-a\cdot\alpha-\lambda-|\alpha|}). \end{aligned}$$

From these estimates and (3.3), we obtain

$$N^{-2+a\cdot\alpha+|\alpha|}I = C(a)C_0^{-a_n-1}\left(\sum_{i,j=1}^{n-1}\partial^a\gamma^{ij}(0)t_it_j + C_0^2\partial^a\gamma^{nn}(0)\right) + O(N^{-\lambda}), \tag{3.4}$$

where $C(a)$ is the quantity defined in (1.14).

Estimates of II and III. We prove that

$$|II| + |III| \le CN^{2-m-|\alpha|-\lambda}. \tag{3.5}$$

Once (3.5) is proved, then Theorem 1.1 follows from (3.2), (3.4), and (3.5).

We only give the proof of (3.5) for II. Equation (3.5) for III can be proved in the same way. Note that

$$\begin{aligned}
II &= \int_\Omega (\gamma\nabla_x s_N)\cdot\overline{\nabla(\zeta_N\Psi_N)}dx\\
&= \int_{\Omega_N\setminus\Omega'_N}(\gamma\nabla_x s_N)\cdot\nabla_x\overline{\Psi_N}dx + \int_{\Omega'_N}(\gamma\nabla_x s_N)\cdot\nabla_x(\overline{\zeta_N\Psi_N})dx\\
&= \int_{\Omega_N\setminus\Omega'_N}(\gamma\nabla_x s_N)\cdot\nabla_x\overline{\Phi_N}dx + \int_{\Omega_N\setminus\Omega'_N}(\gamma\nabla_x s_N)\cdot\nabla_x\overline{(\Psi_N-\Phi_N)}dx\\
&\quad + \int_{\Omega'_N}(\gamma\nabla_x s_N)\cdot\nabla_x(\overline{\zeta_N\Psi_N})dx\\
&:= II_1 + II_2 + II_3.
\end{aligned}$$

In the same way as for (3.3), one can show that

$$|II_3| \le C\exp\left(-\frac{C_0}{4}N^{\frac12}\right). \tag{3.6}$$

We now estimate II_1. Set $D_N := \Omega_N\setminus\Omega'_N$ for convenience. By the definition (3.1) of s_N and Lemma 2.2, we have

$$\begin{aligned}
\|s_N\|_{H^1(D_N)} &\le \|\Phi_N\|_{H^1(D_N)} + \|u_N\|_{H^1(D_N)}\\
&\le \|\Phi_N\|_{H^1(D_N)} + \|\phi_N\|_{H^{1/2}(\partial\Omega)}\\
&\le CN^{1-\frac{|\alpha|}{2}}.
\end{aligned} \tag{3.7}$$

Put $\Gamma_1 := \{|x_j| = N^{-\alpha_j} \text{ for some } j\ (j=1,\dots,n-1),\ 0\le x_n\le\frac{1}{2\sqrt{N}}\}$ and $\Gamma_2 := \{|x_j|\le N^{-\alpha_j}\ (j=1,\dots,n-1),\ x_n=\frac{1}{2\sqrt{N}}\}$. Then $\partial D_N = \Gamma\cup\Gamma_1\cup\Gamma_2$. Since

$$s_N\left(x',\frac{1}{2\sqrt{N}}\right) = \int_0^{\frac{1}{2\sqrt{N}}}\partial_{x_n}s_N(x',t)dt,$$

it follows from the Cauchy–Schwarz inequality and (3.7) that

$$\|s_N\|_{L^2(\Gamma_2)} \le \left(\frac{1}{2\sqrt{N}}\right)^{1/2}\|s_N\|_{H^1(D_N)} \le CN^{\frac34-\frac{|\alpha|}{2}}. \tag{3.8}$$

Observe that $\nabla_x \Phi_N|_{\Gamma_1} = 0$ and $\nabla_x \Phi_N|_{\Gamma_2} = O(e^{-\frac{1}{2}C_0 N^{1/2}})$. Since $s_N = 0$ on Γ, it follows from the divergence theorem that

$$\begin{aligned} II_1 &= \int_{D_N} (\gamma \nabla_x s_N) \cdot \nabla \overline{\Phi_N} dx \\ &= -\int_{D_N} s_N \nabla_x \left(\gamma \nabla_x \overline{\Phi_N}\right) dx + O(e^{-\frac{1}{4}C_0 N^{1/2}}). \end{aligned}$$

Note that

$$\left| \int_{D_N} s_N \nabla_x \left(\gamma \nabla_x \overline{\Phi_N}\right) dx \right| \leq \|x_n \nabla_x (\gamma \nabla_x \overline{\Phi_N})\|_{L^2(D_N)} \|x_n^{-1} s_N\|_{L^2(D_N)}.$$

By the Hardy inequality and (3.7), we have

$$\|x_n^{-1} s_N\|_{L^2(D_N)} \leq C \|s_N\|_{H^1(D_N)} \leq C N^{1-\frac{|\alpha|}{2}}.$$

On the other hand, we obtain from (2.3) that

$$\begin{aligned} \|x_n \nabla_x (\gamma \nabla_x \overline{\Phi_N})\|_{L^2(D_N)} &\leq C N^{2-m-\lambda} \|x_n p(y_n) e^{-C_0 y_n}\|_{L^2(D_N)} \\ &\leq C N^{1-m-\lambda-\frac{|\alpha|}{2}}. \end{aligned} \tag{3.9}$$

It thus follows that

$$|II_1| \leq C N^{2-m-|\alpha|-\lambda}.$$

We now estimate II_2. Note that

$$(\Phi_N - \Psi_N)|_{\Gamma \cup \Gamma_1} = 0, \qquad (\Phi_N - \Psi_N)|_{\Gamma_2} = O\left(e^{-\frac{1}{2}C_0 N^{1/2}}\right).$$

Hence, an integration by parts yields

$$II_2 = -\int_{D_N} \nabla_x (\gamma \nabla_x s_N)(\Psi_N - \Phi_N) dx + O(e^{-\frac{1}{4}C_0 N^{1/2}}).$$

Since $\nabla \cdot (\gamma \nabla u_N) = 0$, we have

$$II_2 = \int_{D_N} \nabla_x (\gamma \nabla_x \Phi_N)(\Psi_N - \Phi_N) dx + O(e^{-\frac{1}{4}C_0 N^{1/2}}).$$

In the same way as for II_1, one can show that

$$|II_2| \leq C N^{2-m-|\alpha|-\lambda}.$$

This completes the proof of (3.5).

4. Proof of Theorem 1.3. In this section we prove Theorem 1.3. Suppose that Ω and g are of the form (1.2) and (1.22). For such a metric g and a multi-index a with $|a| \leq m$, define $g^{a,z}$ to be a positive definite symmetric matrix-valued smooth function on Ω such that

$$g^{a,z}(x) := \sum_{b<a} \frac{\partial^b g(z)}{b!} (x-z)^b \tag{4.1}$$

near z. We then define $\Lambda_{g^{a,z}}$ by

$$\langle \Lambda_{g^{a,z}} f, h\rangle = \int_\Omega (|g|^{-1/2} g^{a,z}\nabla u)\cdot \nabla v dx, \quad f,h\in H^{1/2}(\partial\Omega), \tag{4.2}$$

where $u\in H^1(\Omega)$ is the solution to the problem

$$\begin{aligned} \nabla\cdot(|g|^{-1/2}g^{a,z}\nabla u) &= 0 \quad \text{in } \Omega,\\ u &= f \quad \text{on } \partial\Omega, \end{aligned}$$

and $v\in H^1(\Omega)$ is such that $v|_{\partial\Omega}=h$. If $a=0$, let $\Lambda_{g_0}=0$. Note that $\Lambda_{g^{a,z}}$ is *not* a DtN map corresponding to an invariant Laplacian. Consider it as a DtN map corresponding to a divergence equation.

THEOREM 4.1. *Suppose that $g\in C^{m,p}(\overline{\Omega}\cap B_\delta(0))$. For $z=(z',0)\in\partial\Omega\cap B_\delta(0)$, let*

$$C_g(z) := \sqrt{\sum_{i,j=1}^{n-1} g^{ij}(z)t_i t_j}. \tag{4.3}$$

Then for a multi-index $a=(a',a_n)$ and $k\le m$, we have

$$\begin{aligned} &N^{-2+|\alpha|+a\cdot\alpha}\langle(\Lambda_g-\Lambda_{g^{a,z}})\phi_N^z, \overline{\phi_N^z}\rangle \\ &\quad = C(a)C_g(z)^{-a_n-1}|g(z)|^{-1/2}\sum_{i,j=1}^{n-1}\partial^a g^{ij}(z)t_it_j + O(N^{-\lambda}), \end{aligned} \tag{4.4}$$

where $O(N^{-\lambda})$ is independent of z. In particular, when $a=0$, $C(0)=\frac{1}{2}$, and hence we have

$$N^{-2+|\alpha|}\langle\Lambda_g\phi_N^z,\overline{\phi_N^z}\rangle = 2\sqrt{|g(z)|^{-1}\sum_{i,j=1}^{n-1} g^{ij}(z)t_it_j} + O(N^{-\lambda}). \tag{4.5}$$

Despite a slight difference between Theorem 4.1 and Theorem 1.1, it can be proved in the same way, and so we omit the proof. We are now ready to prove Theorem 1.3.

Proof of Theorem 1.3. Suppose first that g_1 and g_2 are of the forms (1.22) ($j=1,2$) and Ω is of the form (1.2). Let K be a compact subset of Γ such that $\operatorname{dist}(K,\partial\Gamma)>\delta_0$ for some $\delta_0>0$. If N is large enough so that $\operatorname{supp}\phi_N^z\subset\Gamma$ for all $z\in K$, then we have from (2.15) that

$$\begin{aligned} N^{-2+|\alpha|}|\langle(\Lambda_1-\Lambda_2)\phi_N^z,\overline{\phi_N^z}\rangle| &\le N^{-2+|\alpha|}\|\Lambda_1-\Lambda_2\|\|\phi_N^z\|^2_{H^{1/2}(\partial\Omega)}\\ &\le \|\Lambda_1-\Lambda_2\|, \end{aligned}$$

where the norm in the last term is the operator norm from $H^{1/2}(\Gamma)$ into $H^{-1/2}(\Gamma)$. It follows from (4.5) that

$$\left|\sqrt{|g_1(z)|^{-1}\sum_{i,j=1}^{n-1} g_1^{ij}(z)t_it_j} - \sqrt{|g_2(z)|^{-1}\sum_{i,j=1}^{n-1} g_2^{ij}(z)t_it_j}\right| \le C\|\Lambda_1-\Lambda_2\| + O(N^{-\lambda}).$$

Since t' is arbitrary and g_j satisfies (1.20), we have

$$||g_1(z)|^{-1}g_1(z) - |g_2(z)|^{-1}g_2(z)| \le C\|\Lambda_1 - \Lambda_2\| + O(N^{-\lambda}). \tag{4.6}$$

Then, by taking determinants, we have

$$||g_1(z)|^{2-n} - |g_2(z)|^{2-n}| \le C\|\Lambda_1 - \Lambda_2\| + O(N^{-\lambda}),$$

and hence

$$||g_1(z)| - |g_2(z)|| \le C\|\Lambda_1 - \Lambda_2\| + O(N^{-\lambda}). \tag{4.7}$$

It then follows from (4.6) and (4.7) that

$$|g_1(z) - g_2(z)| \le C\|\Lambda_1 - \Lambda_2\|. \tag{4.8}$$

Suppose now that a is a multi-index and $|a| > 0$. Note that

$$\begin{aligned}\langle(\Lambda_1 - \Lambda_2)\phi_N^z, \overline{\phi_N^z}\rangle = \langle(\Lambda_1 - \Lambda_{g_1^{a,z}})\phi_N^z, \overline{\phi_N^z}\rangle - \langle(\Lambda_2 - \Lambda_{g_2^{a,z}})\phi_N^z, \overline{\phi_N^z}\rangle \\ + \langle(\Lambda_{g_1^{a,z}} - \Lambda_{g_2^{a,z}})\phi_N^z, \overline{\phi_N^z}\rangle.\end{aligned}$$

Thus it follows from (4.4) that

$$\begin{aligned}&\left|\frac{\sum_{i,j=1}^{n-1}\partial^a g_1^{ij}(z)t_it_j}{C_{g_1}(z)^{a_n+1}|g_1(z)|^{1/2}} - \frac{\sum_{i,j=1}^{n-1}\partial^a g_2^{ij}(z)t_it_j}{C_{g_2}(z)^{a_n+1}|g_2(z)|^{1/2}}\right| \\ &\le CN^{-2+|\alpha|+a\cdot\alpha}\|\Lambda_1 - \Lambda_2\|\|\phi_N^z\|^2_{H^{1/2}(\partial\Omega)} \\ &\quad + CN^{-2+|\alpha|+a\cdot\alpha}|\langle(\Lambda_{g_1^{a,z}} - \Lambda_{g_2^{a,z}})\phi_N^z, \overline{\phi_N^z}\rangle| + CN^{-\lambda}.\end{aligned} \tag{4.9}$$

Let Φ_N^j be approximate solutions of $\nabla\cdot(|g_j|^{-1/2}g_j^{a,z}\nabla u) = 0$ with the boundary value ϕ_N^z on $\partial\Omega$. Then in the same way as the proof of Theorem 1.1, we can show that

$$\langle(\Lambda_{g_1^{a,z}} - \Lambda_{g_2^{a,z}})\phi_N^z, \overline{\phi_N^z}\rangle = \int_{D_N}|g|^{-1/2}(g_1^{a,z} - g_2^{a,z})\nabla\Phi_N^1\cdot\nabla\overline{\Phi_N^2}dx + O(N^{2-m-|\alpha|-\lambda}).$$

Since

$$g_1^{a,z} - g_2^{a,z} = |g_1|^{-1/2}\sum_{b<a}\frac{\partial^b g_1(z)}{b!}(x-z)^b - |g_2|^{-1/2}\sum_{b<a}\frac{\partial^b g_2(z)}{b!}(x-z)^b,$$

we have

$$\begin{aligned}&\left|\int_{D_N}(g_1^{a,z} - g_2^{a,z})\nabla\Phi_N^1\cdot\nabla\overline{\Phi_N^2}dx\right| \\ &\le C\left||g_1|^{-1/2} - |g_2|^{-1/2}\right|\|\nabla\Phi_N^1\|_{L^2(D_N)}\|\nabla\Phi_N^2\|_{L^2(D_N)} \\ &\quad + C\sum_{b<a}\left|\partial^b g_1(z) - \partial^b g_2(z)\right|\left|\int_{D_N}(x-z)^b\nabla\Phi_N^1\cdot\nabla\overline{\Phi_N^2}dx\right|.\end{aligned}$$

By (2.15) and (2.16), we have

$$\begin{aligned}|\langle(\Lambda_{g_1^{a,z}} - \Lambda_{g_2^{a,z}})\phi_N^z, \overline{\phi_N^z}\rangle| \le C(\|\Lambda_1 - \Lambda_2\| + O(N^{-\lambda}))N^{2-|\alpha|} \\ + C\sum_{b<a}|\partial^b g_1(z) - \partial^b g_2(z)|N^{2-|\alpha|-b\cdot\alpha} + CN^{2-m-|\alpha|-\lambda}.\end{aligned}$$

It then follows from (4.9) and the above estimates that

$$\left|\frac{\sum_{i,j=1}^{n-1}\partial^a g_1^{ij}(z)t_it_j}{C_{g_1}(z)^{a_n+1}|g_1(z)|^{1/2}}-\frac{\sum_{i,j=1}^{n-1}\partial^a g_2^{ij}(z)t_it_j}{C_{g_2}(z)^{a_n+1}|g_2(z)|^{1/2}}\right|$$
$$\leq C\left(\|\Lambda_1-\Lambda_2\|N^{a\cdot\alpha}+\sum_{b<a}|\partial^b g_1(z)-\partial^b g_2(z)|N^{(a-b)\cdot\alpha}+N^{-\lambda}\right).$$

It then follows from (4.8) that

$$|\partial^a g_1(z)-\partial^a g_2(z)| \tag{4.10}$$
$$\leq C\left(\|\Lambda_1-\Lambda_2\|N^{a\cdot\alpha}+\sum_{b<a}|\partial^b g_1(z)-\partial^b g_2(z)|N^{(a-b)\cdot\alpha}+N^{-\lambda}\right).$$

From (4.10), one can show that there exists $C=C(m,\lambda)$ such that

$$|\partial^a g_1(z)-\partial^a g_2(z)|\leq C\|\Lambda_1-\Lambda_2\|^{2^{-a\cdot\alpha/\lambda}}. \tag{4.11}$$

We will give a proof of (4.11) at the end of this paper.

If $|a|=k$, then $a\cdot\alpha\leq k$, and hence we have the following stability: If K is a subset of Γ such that $\operatorname{dist}(K,\Gamma)>\delta_0$ for some $\delta_0>0$, then we have

$$\|g_1-g_2\|_{C_E^k(K)}\leq C\|\Lambda_1-\Lambda_2\|^{2^{-k/\lambda}}. \tag{4.12}$$

We now deal with the general case. Suppose that Γ is an open portion of $\partial\Omega$ and K is a compact subset of Γ. For each $x\in K$, there exists an open neighborhood U_x of x and a diffeomorphism (boundary normal coordinates) $\Phi_{j,x}$ on $U_x\cap\overline{\Omega}_N$ such that $\Phi_{j,x}(U_x\cap\overline{\Omega}_N)$ is of the form (1.2) and $(\Phi_{j,x}^{-1})^*g_j$ is of the form (1.22). Moreover $\Phi_{1,x}(z)=\Phi_{2,x}(z)$ for all $z\in U_x\cap\partial\Omega$. Let K_x be a relatively compact subset of $U_x\cap\Gamma$. Then by (4.8) we have

$$\|(\Phi_{1,x}^{-1})^*g_1-(\Phi_{2,x}^{-1})^*g_2\|_{C_E^k(K_x)}\leq C\|\Lambda_{(\Phi_{1,x}^{-1})^*g_1}-\Lambda_{(\Phi_{2,x}^{-1})^*g_2}\|^{2^{-k/\lambda}}.$$

Put $\phi_x(z):=\Phi_{1,x}(z)=\Phi_{2,x}(z)$ for $z\in U_x\cap\partial\Omega$. Then

$$\Lambda_{(\Phi_{j,x}^{-1})^*g_j}=(\phi_x)_*\Lambda_{g_j}.$$

For the proof of this relation, see [15]. Therefore, we have

$$\|g_1-(\Phi_{2,x}^{-1}\circ\Phi_{1,x})^*g_2\|_{C_E^k(K_x)}\leq C\|(\phi_x)_*(\Lambda_{g_1}-\Lambda_{g_2})\|^{2^{-k/\lambda}}.$$

Put $\Psi_x:=\Phi_{2,x}^{-1}\circ\Phi_{1,x}$. Then we have

$$\|g_1-(\Psi_x)^*g_2\|_{C_E^k(K_x)}\leq C\|\Lambda_{g_1}-\Lambda_{g_2}\|^{2^{-k/\lambda}}.$$

By using a partition of unity, we have the theorem. Note that $\Psi_x=Id$ on $U_x\cap\partial\Omega$ for each x. This completes the proof. □

Derivation of (4.11). Since two multi-indices a,b satisfy $b<a$ if and only if $b\cdot\alpha<a\cdot\alpha$, and $b\cdot\alpha=k\lambda$ for some integer k, we may assign a one-to-one relation from multi-indices a with $|a|\leq m$ into the set $\{0,1,\dots,m/\lambda\}$. Let

$$a(k):=|\partial^b g_1(z)-\partial^b g_2(z)|$$

if $b \cdot \alpha = k\lambda$. Put $a(0) := \|\Lambda_1 - \Lambda_2\|$. Then (4.10) reads

$$a(l) \le C\left(\sum_{k<l} a(k)(N^\lambda)^{l-k} + N^{-\lambda}\right) \quad \text{for all large } N.$$

From this one can show inductively that

$$a(l) \le Ca(0)^{2^{-l}}.$$

Thus (4.11) is obtained. □

Acknowledgments. We would like to express our gratitude to the referees, who read the paper thoroughly and offered several invaluable suggestions.

REFERENCES

[1] G. Alessandrini, *Singular solutions of elliptic equations and the determination of conductivity by boundary measurements*, J. Differential Equations, 84 (1990), pp. 252–273.
[2] G. Alessandrini and R. Gaburro, *Determining conductivity with special anisotropy by boundary measurements*, SIAM J. Math. Anal., 33 (2001), pp. 153–171.
[3] R. M. Brown, *Recovering conductivity at the boundary from the Dirichlet to Neumann map: A pointwise result*, J. Inverse Ill-Posed Probl., 9 (2001), pp. 567–574.
[4] R. M. Brown and G. Uhlmann, *Uniqueness in the inverse conductivity problem for nonsmooth conductivities in two dimensions*, Comm. Partial Differential Equations, 22 (1997), pp. 1009–1027.
[5] V. Isakov, *On uniqueness of recovery of a discontinuous conductivity coefficient*, Comm. Pure Appl. Math., 41 (1988), pp. 865–877.
[6] R. Kohn and M. Vogelius, *Determining conductivity by boundary measurements*, Comm. Pure Appl. Math., 37 (1984), pp. 281–298.
[7] R. Kohn and M. Vogelius, *Determining conductivity by boundary measurements* II, *Interior results*, Comm. Pure Appl. Math., 38 (1985), pp. 643–667.
[8] M. Lassas and G. Uhlmann, *On determining a Riemannian manifold from the Dirichlet-to-Neumann map*, Ann. Sci. École Norm. Sup. (4), 34 (2001), pp. 771–787.
[9] J. Lee and G. Uhlmann, *Determining anistropic real-analytic conductivities by boundary measurements*, Comm. Pure Appl. Math., 42 (1989), pp. 1097–1112.
[10] W. R. B. Lionheart, *Conformal uniqueness results in anisotropic electrical impedance imaging*, Inverse Problems, 13 (1997), pp. 125–134
[11] A. Nachman, *The inverse reconstructions from boundary measurements*, Ann. of Math., 128 (1988), pp. 531–587.
[12] A. Nachman, *Global uniqueness for a two-dimensional inverse boundary value problem*, Ann. of Math., 143 (1996), pp. 71–96.
[13] G. Nakamura and K. Tanuma, *Local determination of conductivity at the boundary from Dirichlet to Neumann map*, Inverse Problems, 17 (2001), pp. 405–419.
[14] G. Nakamura and K. Tanuma, *Direct determination of the derivatives of conductivity at the boundary from the localized Dirichlet to Neumann map*, Second Japan-Korea Joint Seminar on Inverse Problems and Related Topics (Seoul, 2001), Commun. Korean Math. Soc., 16 (2001), pp. 415–425.
[15] J. Sylvester, *An anisotropic inverse boundary value problem*, Comm. Pure Appl. Math., 43 (1990), pp. 201–232.
[16] J. Sylvester and G. Uhlmann, *A uniqueness theorem for an inverse boundary value problem in electrical prospection*, Comm. Pure Appl. Math., 39 (1986), pp. 91–112.
[17] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math., 125 (1987), pp. 153–169.
[18] J. Sylvester and G. Uhlmann, *Inverse boundary value problems at the boundary-continuous dependence*, Comm. Pure Appl. Math., 41 (1988), pp. 197–221.

SIAM J. MATH. ANAL.
Vol. 34, No. 3, pp. 736–758

FUNCTIONAL EQUATIONS AND POINCARÉ INVARIANT MECHANICAL SYSTEMS*

J. G. B. BYATT-SMITH† AND H. W. BRADEN†

Abstract. We study the following functional equation that has arisen in the context of mechanical systems invariant under the Poincaré algebra:

$$\sum_{i=1}^{n+1} \frac{\partial}{\partial x_i} \prod_{j \neq i} f(x_i - x_j) = 0, \qquad n \geq 2.$$

New techniques are developed and the general solution within a certain class of functions is given. New solutions are found.

Key words. functional equation, mechanics

AMS subject classifications. 39B32, 33E05

PII. S0036141001393742

1. Introduction.

The differential equation

$$\begin{vmatrix} 1 & 1 & 1 \\ f(u) & f(v) & f(w) \\ f'(u) & f'(v) & f'(w) \end{vmatrix} = 0 \quad \text{subject to } u + v + w = 0 \tag{1.1}$$

is one form of the first of a series of differential equations that can be written as

$$\sum_{i=1}^{n+1} \frac{\partial}{\partial x_i} \prod_{j \neq i} f(x_i - x_j) = 0, \qquad n \geq 2, \tag{1.2}$$

subject to the constraint that f is an even function. Equation (1.1) appeared in the context of a three-body integrable quantum mechanical problem and was studied by Buchstaber and Perelomov [7] and later by Braden and Byatt-Smith [2]. Equations (1.2), which have appeared in the context of constructing mechanical systems with certain invariances, are the focus of this paper. (We will later review these connections between functional equations and mechanical systems.) When $n = 2$, (1.2) gives

$$\begin{aligned} &\frac{\partial}{\partial x_1}\left(f(x_1 - x_2) f(x_1 - x_3)\right) + \frac{\partial}{\partial x_2}\left(f(x_2 - x_3) f(x_2 - x_1)\right) \\ &\qquad + \frac{\partial}{\partial x_3}\left(f(x_3 - x_1) f(x_3 - x_2)\right) = 0. \end{aligned} \tag{1.3}$$

Using the evenness of f we may express this as

$$\begin{aligned} &\frac{\partial}{\partial x_1}\left(f(x_1 - x_2) f(x_3 - x_1)\right) + \frac{\partial}{\partial x_2}\left(f(x_2 - x_3) f(x_1 - x_2)\right) \\ &\qquad + \frac{\partial}{\partial x_3}\left(f(x_3 - x_1) f(x_2 - x_3)\right) = 0. \end{aligned} \tag{1.4}$$

*Received by the editors August 14, 2001; accepted for publication (in revised form) July 24, 2002; published electronically January 28, 2003.
http://www.siam.org/journals/sima/34-3/39374.html
†Department of Mathematics and Statistics, The University of Edinburgh, Edinburgh, UK (Byatt @ed.ac.uk, hwb@ed.ac.uk).

Equation (1.1) can be written in this form if we put $u = x_1 - x_2, v = x_2 - x_3$, and $w = x_3 - x_1$, which automatically satisfies the constraint, although the assumption of evenness is not required.

Braden and Byatt-Smith [2] proved, as part of a more general theorem, that the complete solution set for the function f which satisfies (1.1) is

$$f(u) = a + bu \text{ or } a + b\, e^{cu} \tag{1.5a}$$

or

$$f(u) = a + b\,\wp(cu + d, g_2, g_3)\,. \tag{1.5b}$$

Here $\wp$ is the Weierstrass $\wp$-function. Equation (1.5b) has six constants associated with it, namely, $\{a, b, c, d, g_2, g_3\}$, where g_2 and g_3 are the two constants which relate to the two periods of $\wp(z)$, and d is one-third of any period. However, $\wp(z)$ satisfies the equation $\wp'^2 = 4\wp^3 - g_2\wp - g_3$, with $z^2\wp(z) \to 1$ as $z \to 0$. Hence $\wp$ satisfies the scaling law $\wp(cu, g_2, g_3) \equiv c^{-2}\wp\left(u, c^4 g_2, c^6 g_3\right)$, so that without loss of generality we may take $c = 1$ in (1.5b). The solution set represented by (1.5a), which is a subset of (1.5b), does not require the constraint $u + v + w = 0$ to be satisfied and is the general solution of the differential equation

$$f' f''' - f''^2 = 0. \tag{1.6}$$

The solution set (1.5b) only satisfies (1.1) provided the constraint is satisfied and is the general solution of

$$f'^2 + Af^3 + Bf^2 + Cf + D = 0, \tag{1.7}$$

or, upon eliminating the arbitrary constants A, B, C, and D,

$$f'^2 f^{(v)} - 3f' f'' f^{(iv)} + 3f''^2 f''' - f' f'''^2 = 0. \tag{1.8}$$

When we consider (1.5b) as the general solution of (1.7) or (1.8), d appears as an arbitrary constant. However, if (1.5b) is also to satisfy (1.1), then d is not arbitrary and, by substituting the solution back into (1.1), can be shown to be either zero or any integer multiple of one-third of any period of $\wp(z)$. A simpler proof that all of the solutions of (1.1) are contained in the solution set of (1.5a, b), subject to the above condition on d, can be obtained by eliminating the functions $f(v)$ and $f(w)$ by taking suitable combinations of derivatives of (1.1). At various points in the proof the equation factorizes to give (1.6) or (1.7) as factors. (This was the approach of [2].)

For the last 15 years the nature of the solutions to (1.2) has remained open. One can show [13] that (1.5b), with $d = 0$ to ensure the evenness of f, satisfy (1.2), and it has been conjectured that such were the only solutions. Though the method of eliminating functions just noted above for the case of (1.1) should be applicable in the general case, the algebra involved to completely define the solution set is quite considerable. In the case of (1.2), when $n \geq 3$ the amount of algebra appears to be so large that even a Maple calculation cannot handle the details necessary to provide a definition of the solution set. Here we shall develop new techniques to handle the equation. Our main result contains a surprise. To describe this let us introduce the following.

DEFINITION 1. *Let $\mathcal{A}$ and $\mathcal{A}'$ denote the class of meromorphic functions $f(z)$ with the following properties:*

(1) *$f(z)$ is an even function of z.*

(2) *The only singularity of $f(z)$ in the strip $|\mathcal{I}m(z)| < a$ for some $a > 0$ is a double pole at the origin. And (respectively)*

(3) *$\frac{z^2 f(z)}{2a^2+z^2}$ is L_1 along any line $|\mathcal{I}m(z)| = t,\ -\infty < \mathcal{R}e(z) < \infty$, where $|t| < a$.*

(3′) (a) *Either $z^2 f(z)$ is L_1 along any line $|\mathcal{I}m(z)| = t,\ -\infty < \mathcal{R}e(z) < \infty$, where $|t| < a$, or* (b) *there exists a constant c such that $z^2 f(z) - c$ is L_1 along any line $|\mathcal{I}m(z)| = t,\ -\infty < |\mathcal{R}e(z) < \infty$, where $|t| < a$.*

DEFINITION 2. *Let $\mathcal{A}_p$ denote the class of periodic, meromorphic functions $f(z)$ whose period is p (where p is real) with the following properties:*

(1) *$f(z)$ is an even function of z.*

(2) *The only singularities of $f(z)$ in the strip $|\mathcal{I}m(z)| < a$ for some $a > 0$ are a periodic array of double poles of the form $1/(z - 2n\pi)^2$ at the points $z = 2n\pi$, $n \in \mathbb{Z}$, on the real axis.*

With these definitions in hand we find the following.

THEOREM 1. *For functions $f(z) \in \mathcal{A}' \cup \mathcal{A}_p$ the general even solution of* (1.2) *is*

(a) *for all even $n \geq 2$ given by* (1.5b) *with $d = 0$, while*

(b) *for odd $n \geq 3$ there are in addition to the solutions* (1.5b) *with $d = 0$ the following:*

$$h_1(z) = \sqrt{(\wp(z) - e_2)(\wp(z) - e_3)} = \frac{\sigma_2(z)\sigma_3(z)}{\sigma^2(z)} = \frac{\theta_3(v)\theta_4(v)}{\theta_1^2(v)} \frac{\theta_1'^2(0)}{4\omega^2\theta_3(0)\theta_4(0)} = b\frac{\mathrm{dn}(u)}{\mathrm{sn}^2(u)},$$

$$h_2(z) = \sqrt{(\wp(z) - e_1)(\wp(z) - e_3)} = \frac{\sigma_1(z)\sigma_3(z)}{\sigma^2(z)} = \frac{\theta_2(v)\theta_4(v)}{\theta_1^2(v)} \frac{\theta_1'^2(0)}{4\omega^2\theta_2(0)\theta_4(0)} = b\frac{\mathrm{cn}(u)}{\mathrm{sn}^2(u)},$$

$$h_3(z) = \sqrt{(\wp(z) - e_1)(\wp(z) - e_2)} = \frac{\sigma_1(z)\sigma_2(z)}{\sigma^2(z)} = \frac{\theta_2(v)\theta_3(v)}{\theta_1^2(v)} \frac{\theta_1'^2(0)}{4\omega^2\theta_2(0)\theta_3(0)} = b\frac{\mathrm{cn}(u)\,\mathrm{dn}(u)}{\mathrm{sn}^2(u)}.$$

Here

$$\sigma_\alpha(z) = \frac{\sigma(z + \omega_\alpha)}{\sigma(\omega_\alpha)} e^{-z\zeta(\omega_\alpha)}, \qquad u = \sqrt{e_1 - e_3}\, z, \qquad v = \frac{z}{2\omega}, \qquad b = e_1 - e_3,$$

with $\omega_1 = \omega$, $\omega_2 = -\omega - \omega'$, and $\omega_3 = \omega'$, and we have given representations in terms of the Weierstrass elliptic functions, theta functions, and the Jacobi elliptic functions [16]. For appropriate ranges of z the solutions are real. These exhaust the even periodic solutions of (1.2), and their degenerations yield all the even solutions with only a double pole at $x = 0$ on the real axis. The assumption of only a double pole at the origin comes from an elementary singularity analysis of (1.2). When $n = 2$, 3 the theorem can be proved without the assumption that $f(z) \in \mathcal{A}' \cup \mathcal{A}_p$, provided it is assumed that f is meromorphic with a double pole at the origin. We also conjecture this latter assumption is all that is required for $n \geq 4$ but have been unable to prove this. The surprise is the appearance of these new solutions for odd n, which in turn yield new Poincaré invariant mechanical systems.

Our paper is arranged as follows. First we will describe the origin of (1.2) and of the several connections between functional equations and mechanical systems in section 2. Section 3 is the heart of the paper where we introduce our new method. Here we take a (suitable) Fourier transform of (1.2), turning that functional equation into a functional equation for the transform. It is in taking this Fourier transform that we encounter the class $\mathcal{A}$, which is sufficient for the transform to exist. The

functional equation we produce has a remarkable property: we can reduce the general n equation to a consideration of the $n = 2$ and $n = 3$ cases. Subject to one or two assertions proven in sections 4 and 5, we have essentially proven the theorem. Section 4 considers the case of the $n = 2$ equation and section 5 considers the case of the $n = 3$, which yields the new solutions. It is only at this stage in solving these equations do we need the more restricted class $\mathcal{A}'$. At this stage we have proven our theorem, and the remaining two sections are included for completeness. In section 6 we consider using our *Fourier transform method* to rederive the general solution of (1.1) found in [2]. Finally in section 7 we consider the more common series methods for solving functional equations, as used, for example, in [2]. These methods yield either a Laurent series for the solution set or a set of differential equations, whose common solution the solution set must satisfy. The advantage of these methods is that they define the solution set, or the differential equations that the set must satisfy, for (1.2) when $n = 2$ or $n = 3$. Even though these methods appear intractable for larger values of n, the equations derived for $n = 2$ and 3 can be solved, and we can prove theorem 1 for cases $n = 2$ and 3 only, but under weaker conditions than those required for the *Fourier transform method*.

2. Some mechanical systems. Some years ago Ruijsenaars and Schneider [13] initiated the study of mechanical systems exhibiting an action of Poincaré algebra:

$$\{H, B\} = P, \qquad \{P, B\} = H, \qquad \{H, P\} = 0. \tag{2.1}$$

Here H is the Hamiltonian of the system generating time-translations, P is a space-translation generator, and B is the generator of boosts. The models they discovered were found to posses other nice features: they were in fact integrable and a quantum version of them naturally existed. These models also appear in various field theoretic contexts (see [5]). Ruijsenaars and Schneider began with the ansatz for a system of $n + 1$ particles interacting on the line,

$$H = \sum_{j=1}^{n+1} \cosh p_j \prod_{k \neq j} F(x_j - x_k), \qquad P = \sum_{j=1}^{n+1} \sinh p_j \prod_{k \neq j} F(x_j - x_k),$$

and

$$B = -\sum_{j=1}^{n+1} x_j.$$

With this ansatz and the canonical Poisson bracket $\{x_i, p_j\} = \delta_{ij}$ the first two Poisson brackets of (2.1) involving the boost operator B are automatically satisfied. Supposing further that $F(x) = \pm F(-x)$, then the final Poisson bracket is equivalent to the functional equation

$$\{H, P\} = 0 \Longleftrightarrow \sum_{j=1}^{n+1} \partial_j \prod_{k \neq j} F^2(x_j - x_k) = 0. \tag{2.2}$$

With $f(x) = F(x)\,F(-x)$ this is precisely (1.2), and so solutions of this equation yield Poincaré invariant mechanical systems. At the time, Ruijsenaars and Schneider were able to show that (1.5b) (with $d = 0$) gave solutions to these equations for all n. These solutions in fact yield $n+1$ independent, mutually Poisson commuting conserved

quantities, and so are a completely integrable mechanical system. A scaling limit of the Ruijsenaars–Schneider model yields the Calogero–Moser system with Hamiltonian

$$H = \frac{1}{2}\sum_{i=1}^{n+1} p_i^2 + \frac{1}{6}\sum_{i\neq j} \wp\left(q_i - q_j\right), \tag{2.3}$$

which is another well-studied completely integrable system [15].

We note that many connections exist between functional equations and integrable quantum and classical systems [3, 4, 6, 7, 8, 9, 11, 12]. Functional equation (1.1) (without any assumptions on the parity of the function f) arises, for example, when characterizing quantum mechanical potentials whose ground state wavefunction (of a given form) is factorizable [10, 14]. Buchstaber and Perelomov solved this equation in [7]. More recently, it has been shown to characterize the Calogero–Moser system [1]. Several functional equations appearing in this setting and whose general solutions have still to be found are given in [4].

3. The Fourier transform method. Our strategy to solve (1.2) will be to derive a functional equation for the Fourier transform of this equation, which we then proceed to solve. Indeed, we will show that in order to solve the functional equation for the Fourier transform, it suffices to solve for the $n = 2$ and $n = 3$ cases, and this is done in later sections. In deriving the functional equation for the Fourier transform we need to assume that $f \in \mathcal{A}$. However, in order to derive the solutions to the $n = 2$ and $n = 3$ equations via this Fourier transform method, we need to assume the more restrictive condition $f \in \mathcal{A}'$. (In section 7 we see that this further restriction is unnecessary.)

Let us then look first to the problem of deriving the equation for the Fourier transform of the solution of (1.1) and (1.2) for general integer n. There are a variety of difficulties which are not immediately apparent. These will be treated as they arise. They include requiring an appropriate generalized Fourier transform for functions of x which are unbounded either at infinity or at points on the real axis, and the consideration of distributional solutions both of (1.2), (1.3), and (1.4) and of the transformed equation. We first consider the even solutions of (1.2) for a general integer n. It will be convenient to write (1.2) as

$$g\left(\mathbf{x}, x_{n+1}\right) = \sum_{p=1}^{n+1} \frac{\partial}{\partial x_p} \prod_{q\neq p} f\left(x_p - x_q\right), \tag{3.1}$$

where f is even and $\mathbf{x}$ is the vector $(x_1, x_2, \ldots, x_n)$. We then define the n-dimensional Fourier transform $\widehat{g}(\mathbf{k}, x_{n+1})$ of $g\left(\mathbf{x}, x_{n+1}\right)$ by

$$\widehat{g}\left(\mathbf{k}, x_{n+1}\right) = \int_{\mathbb{R}^n} g\left(\mathbf{x}, x_{n+1}\right) e^{-i\mathbf{k}.\mathbf{x}} d\mathbf{x}. \tag{3.2}$$

However, there are problems with the double pole of f at the origin. Away from these singularities g is identically zero for these solutions. However, to interpret (3.2) correctly we find that g acts as a distribution over any plane through the origin, and the 2-dimensional Fourier transform of g over this plane gives a nonzero contribution. The contribution from all of these singularities becomes difficult to deal with. To overcome this we assume that the arguments of g are complex, and we replace x_j by

$x_j + i\epsilon_j$, where $\epsilon_{n+1} > \epsilon_n > \cdots > \epsilon_1 > 0$, and assume that ϵ_{n+1} is small. In other words

$$g = \sum_{p=1}^{n+1} \frac{\partial}{\partial x_p} \prod_{q \neq p} f\left(x_p - x_q + i\left(\epsilon_p - \epsilon_q\right)\right). \tag{3.3}$$

We then assume that in the definition of $\widehat{g}$, (3.2), we integrate along the real axis in the complex $x_j + iy_j$ plane. If the function $f(z)$ has double poles on the $x = Re\ z$ axis and no other singularities in the neighborhood of the x-axis, then, provided ϵ_{n+1} is small enough, there will be no singularities of g within the domain of integration of the integral occurring in (3.2). This is the motivation for item (2) in Definitions 1 and 2.

To calculate $\widehat{g}$ we make some changes of variables. Set $\xi_q = x_1 - x_q + i\left(\epsilon_1 - \epsilon_q\right)$ for $2 \leq q \leq n+1$ and $\xi_1 = x_1 - x_{n+1} + i\left(\epsilon_1 - \epsilon_{n+1}\right)$. Then $\xi_1 = \xi_{n+1}$. Further set $X_q = x_1 - x_q$. Then focusing first on the $p = 1$ term in (3.2) consider

$$\begin{aligned} I &= \int_{x_1=-\infty}^{\infty} \int_{\mathbb{R}^{n-1}} \prod_{q=2}^{n+1} f\left(\xi_q\right) e^{-i\mathbf{k}.\mathbf{x}} d^{n-1}x\, dx_1 \\ &= \int_{x_1=-\infty}^{\infty} f(\xi_1) e^{-ik_1x_1} \left(\prod_{q=2}^{n} \int_{x_q=-\infty}^{\infty} f\left(\xi_q\right) e^{-ik_qx_q} dx_q \right) dx_1 \\ &= \int_{x_1=-\infty}^{\infty} f(\xi_1) e^{-ik_1x_1} \left(\prod_{q=2}^{n} \int_{X_q=-\infty}^{\infty} f\left(\xi_q\right) e^{ik_q(X_q - x_1)} dX_q \right) dx_1. \end{aligned} \tag{3.4}$$

Now since $\epsilon_q > \epsilon_1$ the singularities of $f\left(z_1\right)$ at $z_1 = i(\epsilon_{n+1} - \epsilon_1)$, and (for $q > 1$) of $f\left(z_q\right)$ at $z_q = i(\epsilon_q - \epsilon_1)$ lie in the upper half plane.

For functions $f(z)$ which are analytic in the neighborhood of the $Re\ z$ axis but have a pole at $z = 0$ we define Fourier transforms $\widehat{f}_U$ and $\widehat{f}_L$ by

$$\widehat{f}_U(k) = \int\!\!\!\!\!\!\uparrow_{-\infty}^{\infty} f(x) e^{-ikx} dx \tag{3.5}$$

and

$$\widehat{f}_L(k) = \int\!\!\!\!\!\!\downarrow_{-\infty}^{\infty} f(x) e^{-ikx} dx, \tag{3.6}$$

which are suitably indented to go above and below the singularity at the origin. With these definitions we have, in the limit as $\epsilon_{n+1} \to 0$,

$$\begin{aligned} I &= \int_{-\infty}^{\infty} f\left(x_1 - x_{n+1} + i\left(\epsilon_1 - \epsilon_{n+1}\right)\right) e^{-i\left(\sum_{q=1}^{n} k_q\right)x_1} dx_1 \prod_{q=2}^{n} \widehat{f}_L(-k_q) \\ &= e^{-i\left(\sum_{q=1}^{n} k_q\right)x_{n+1}} \widehat{f}_L\left(\sum_{q=1}^{n} k_q\right) \prod_{q=2}^{n} \widehat{f}_L(-k_q). \end{aligned} \tag{3.7}$$

Assuming condition (3) of Definition 1, the Fourier transform of f exists and is continuous. Thus the first term in the sum in (3.3) contributes a term ik_1I to $\widehat{g}$. A

similar calculation gives

$$\widehat{g} = \left(\sum_{j=1}^{n}\left[k_j\left(\prod_{q=1}^{j-1}\widehat{f}_U(-k_q)\right)\left(\prod_{q=j+1}^{n}\widehat{f}_L(-k_q)\right)\right]\widehat{f}_L\left(\sum_{q=1}^{n}k_q\right) - \left(\sum_{j=1}^{n}k_j\right)\left(\prod_{q=1}^{n}\widehat{f}_U(-k_q)\right)\right) ie^{-i\left(\sum_{q=1}^{n}k_q\right)x_{n+1}}, \tag{3.8}$$

so that $\widehat{g}=0$ gives

$$\sum_{j=1}^{n}\left[k_i\left(\prod_{q=1}^{j-1}\widehat{f}_U(-k_q)\right)\left(\prod_{q=j+1}^{n}\widehat{f}_L(-k_q)\right)\right]\widehat{f}_L\left(\sum_{q=1}^{n}k_q\right)-\left(\sum_{j=1}^{n}k_j\right)\left(\prod_{q=1}^{n}\widehat{f}_U(-k_q)\right)=0. \tag{3.9}$$

We now define the Fourier transform of $f(z)$ to be $\frac{1}{2}(\widehat{f}_U(k)+\widehat{f}_L(k))$. Also we have the result that the difference $\widehat{f}_L(k)-\widehat{f}_U(k)$ is $2\pi i$ multiplied by the residue of the functions $f(z)\,e^{-ikz}$ at the origin, assuming that the only singularity of $f(z)$ on the real axis is at the origin. We further make the assumption that $f(z)$ is an even function of z with a double pole of the form $1/z^2$ at the origin. Hence with

$$\widehat{f}(k)=\frac{1}{2}\left(\widehat{f}_U(k)+\widehat{f}_L(k)\right) \tag{3.10}$$

and

$$\widehat{f}_L(k)-\widehat{f}_U(k)=2\pi i\times\ \text{Residue}\left(f(z)\,e^{-ikz}\right)=2\pi k, \tag{3.11}$$

we have

$$\widehat{f}_L(k)=\widehat{f}(k)+\pi k \tag{3.12}$$

and

$$\widehat{f}_U(k)=\widehat{f}(k)-\pi k. \tag{3.13}$$

Hence the equation $\widehat{g}=0$, (3.9) gives

$$\begin{aligned} S_n &\equiv \sum_{j=1}^{n}\left[k_j\prod_{q=1}^{j-1}\left(\widehat{f}(-k_q)+\pi k_q\right)\prod_{q=j+1}^{n}\left(\widehat{f}(-k_q)-\pi k_q\right)\right]\left(\widehat{f}\left(\sum_{q=1}^{n}k_q\right)+\pi\sum_{q=1}^{n}k_q\right) \\ &\qquad -\left(\sum_{j=1}^{n}k_j\right)\left(\prod_{q=1}^{n}\left(\widehat{f}(-k_q)+\pi k_q\right)\right) \\ &= \sum_{j=1}^{n}\left[k_j\prod_{q=1}^{j-1}\left(\widehat{f}(k_q)+\pi k_q\right)\prod_{q=j+1}^{n}\left(\widehat{f}(k_q)-\pi k_q\right)\right]\left(\widehat{f}\left(\sum_{q=1}^{n}k_q\right)+\pi\sum_{q=1}^{n}k_q\right) \\ &\qquad -\left(\sum_{j=1}^{n}k_j\right)\left(\prod_{q=1}^{n}\left(\widehat{f}(k_q)+\pi k_q\right)\right)=0, \end{aligned} \tag{3.14}$$

the final expression being obtained using the fact that $\widehat{f}(k)$ is an even function of k. This equation is to be regarded as one which determines $\widehat{f}(k)$ and must be satisfied for all $\{k_q\}$ in $\mathbb{R}^n$. This is the desired functional equation for the Fourier transform of our equation.

We note that, by inspection, $\widehat{f}(k) = \pm\pi|k|$ satisfies (3.14) for all n, and we may also show that $\widehat{f}(k) = t\pi|k|$ satisfies (3.4) for all values of t which satisfy $(1+t)^n(t-1) = (t-1)^n(1+t)$. Apart from $t = \pm 1$, these are $t = i\cot\frac{j\pi}{n-1}$, $j = 1, \ldots, n-2$. However, to satisfy the requirement that the inverse $f(z)$ is such that $z^2 f(z) \to 1$ as $z \to 0$, we require the solution above with $t = -1$. This problem recurs throughout the rest of the paper, and from now on it will be assumed that we take only the multiple of $\widehat{f}(k)$ which satisfies the criterion that it has the correct double pole either at the origin or at the sequence of double poles when $f(z)$ is periodic.

At this stage we have identified a solution to (3.14), the Fourier transform of (1.2). The following lemmas prove useful in finding the complete solution to (3.14) for all n. We begin with a definition.

DEFINITION 3. *For each n let $\mathcal{F}_n$ be the solution set of $S_n = 0$, and let $\mathcal{G}_n$ be the solution set of $S_n = 0$ and $\widehat{f}(0) \neq 0$.*

LEMMA 1. *For $n \geq 4$ and even, $\mathcal{F}_n \subseteq \mathcal{F}_2$.*

LEMMA 2. *For $n \geq 3$, $\mathcal{G}_n \subseteq \mathcal{G}_2$.*

Proof of Lemma 1. In S_{n+2} we put $k_{n+2} = -k_{n+1}$, and using the fact that $\widehat{f}$ is even, we find

$$S_{n+2}\Big|_{k_{n+2}=-k_{n+1}} = \left(\widehat{f}^2(k_{n+1}) - \pi^2 k_{n+1}^2\right) S_n. \tag{3.15}$$

The factor $\widehat{f}^2(k) - \pi^2 k^2$ produces the solutions $\widehat{f}(k) = \pm\pi|k|$ if $\widehat{f}$ is even. This function we have already shown belongs to all the solution sets $\mathcal{F}_n$. Since the solution set $\mathcal{F}_{n+2}$ must be contained in the solution set of $S_{n+2}\big|_{k_{n+2}=-k_{n+1}}$, (3.15) shows that $\mathcal{F}_{n+2} \subseteq (\widehat{f}(k) = \pm\pi|k|) \cup \mathcal{F}_n \subseteq \mathcal{F}_n$. The result now follows by induction.

Proof of Lemma 2. In S_{n+1} we put $k_{n+1} = 0$ and obtain

$$S_{n+1}\Big|_{k_{n+1}=0} = \widehat{f}(0)\, S_n. \tag{3.16}$$

Again if $\widehat{f}(0) \neq 0$, the result follows by induction.

DEFINITION 4. *Let $\mathcal{B}$ denote the class of functions whose generalized Fourier transform $\widehat{f}(k)$ arise from functions $f(z) \in \mathcal{A}'$.*

Although we have derived (3.14) for the class of functions $\mathcal{A}$, only for the subclass of functions given by $\mathcal{A}'$ are we able to show the following theorem.

THEOREM 2. *For solutions with $\widehat{f} \in \mathcal{B}$, then $\mathcal{F}_n = \mathcal{F}_2$ for all even n.*

Proof. By Lemma 1 $\mathcal{F}_n \subseteq \mathcal{F}_2$. In section 4 we will prove that $\mathcal{F}_2$ is the one parameter family $\widehat{f}(k) = \pi k \coth(\pi k/a_0)$, together with its limit as $a_0 \to 0$, $\pm\pi|k|$. It can be verified by substitution that these satisfy $S_n = 0$, as we show in section 5 (see 5.9). This gives $\mathcal{F}_2 \subseteq \mathcal{F}_n$, and the result now follows.

THEOREM 3. *For solutions with $\widehat{f} \in \mathcal{B}$, then $\mathcal{G}_n = \mathcal{G}_2$ for all n.*

The proof of Theorem 3 follows by the same method as the proof of Theorem 2. However, we cannot exclude the possibility that there exist solutions of $S_n = 0$, with $n \geq 3$ and odd, which have $\widehat{f}(0) = 0$ but are not contained in the solutions of $S_2 = 0$. This will be the way our new solutions arise. At this stage, subject to our assertions regarding the $n = 2$ solutions to be proven, we have reduced the problem of solution of the functional equation of the Fourier transform to the cases of $n = 2$ and $n = 3$

with $\widehat{f}(0) = 0$. Before we look at these cases we conclude this section by considering the extensions of Lemmas 1 and 2 and the corresponding theorems about the solution sets to functions $f(z)$ which are periodic. This requires $f(z)$ to have a periodic array of double poles of the form $1/(z - 2n\pi)^2$ at the points $z = 2n\pi$, $n \in \mathbb{Z}$.

To cater for the even periodic solutions of (1.5b) that are also bounded on the real axis, apart from a periodic sequence of double poles, we assume that the solutions $f(x)$ are 2π-periodic and write

$$f(x) = \frac{1}{2\pi} \sum_{p=-\infty}^{\infty} a_p e^{ipx}, \quad \text{with } a_{-p} = a_p. \tag{3.17}$$

The factor 2π is introduced so that the Fourier transform takes the form

$$\widehat{f}(k) = \sum_{p=-\infty}^{\infty} a_p \delta(k - p). \tag{3.18}$$

The Fourier coefficients for a well-behaved function are defined in the usual way in terms of an integral over a period. The corresponding generalized definition for functions with singularities is defined in the same way as the Fourier transform f_U and f_L via (3.5) and (3.6) by taking the integrals, suitably indented, over a period. The pole contributions to the Fourier series in (3.5) and (3.6) then give rise to a term $\sum_{p=-\infty}^{\infty} \pi k_p \delta(k - p)$, which replaces the corresponding term πk_p in (3.14), using the same interpretation as that in (3.18). We introduce these expressions into (3.14) and then recognize that $\widehat{f}(k)$ is only nonzero when k is an integer. Thus, by replacing $\widehat{f}(k_q)$ by $a_{K_q}\delta(k_q - K_q)$ and the pole contribution by $K_q\delta(k_q - K_q)$, in the neighborhood of the point $\{k_p = K_p,\, p = 1, \ldots, n\}$ we obtain

$$\begin{aligned} S_n = \sum_{j=1}^{n} \Bigg[K_j \prod_{q=1}^{j-1} \left((a_{K_q} + \pi K_q)\delta(k_q - K_q) \right) \prod_{q=j+1}^{n} \left((a_{K_q} - \pi K_q)\delta(k_q - K_q) \right) \Bigg] \\ \times \left(\left(a_{\sum_{q=1}^{n} K_q} + \pi \sum_{q=1}^{n} K_q \right) \delta\left(\sum_{q=1}^{n} (k_q - K_q) \right) \right) \\ - \left(\sum_{j=1}^{n} K_j \right) \left(\prod_{q=1}^{n} \left((a_{K_q} + \pi K_q)\delta(k_q - K_q) \right) \right) = 0. \end{aligned} \tag{3.19}$$

This is a distribution supported at $k_q = K_q$ $(q = 1, \ldots, n)$ and is identically zero if

$$\begin{aligned} \sum_{j=1}^{n} \Bigg[K_j \prod_{q=1}^{j-1} \left(a_{K_q} + \pi K_q \right) \prod_{q=j+1}^{n} \left(a_{K_q} - \pi K_q \right) \Bigg] \left(a_{\sum_{q=1}^{n} K_q} + \pi \sum_{q=1}^{n} K_q \right) \\ - \left(\sum_{j=1}^{n} K_j \right) \left(\prod_{q=1}^{n} \left(a_{K_q} + \pi K_q \right) \right) = 0 \end{aligned} \tag{3.20}$$

for all integer values of K_q. Since $\widehat{f}(k)$ satisfy the continuous version of (3.20), it is clear that the solution $a_K = \widehat{f}(K)$, $K \neq 0$ with a_0 arbitrary, will satisfy (3.20). It is

also easy to construct the corresponding function $f(x)$ that satisfies (1.4). If $f(x)$ is the function whose Fourier transform is $\widehat{f}(k)$, then we define

$$h(x) = \sum_{p=-\infty}^{\infty} f(x - 2\pi p). \tag{3.21}$$

This function is 2π-periodic and has Fourier series $\frac{1}{2\pi}\sum_{K=-\infty}^{\infty} a_K e^{iKx}$, where

$$a_K = \int_0^{2\pi} \sum_{p=-\infty}^{\infty} f(x - 2p\pi)\, e^{-iKx} dx = \int_{-\infty}^{\infty} f(x)\, e^{-iKx} dx = \widehat{f}(K). \tag{3.22}$$

The above solution for a_K can also be obtained directly from (3.20) by successively solving all equations with $1 \le |K_q| \le N$, $q = 1, \ldots, n$, for $N = 1, 2, 3, \ldots$, the equations where $K_q = 0$ being automatically satisfied. All the solutions to (3.20) are obtained in section 4.

The conclusions can be summed up in the following theorem, which incorporates the result of Braden and Byatt-Smith [2] for even functions as a special case.

DEFINITION 5. *Let $\mathcal{B}_{2\pi}$ denote the class of functions whose generalized Fourier transform $\widehat{f}(k)$ arises from 2π-periodic meromorphic functions $f(z)$, bounded on the real axis apart from double poles at $1/(z - 2\pi n)^2$, $n \in \mathbb{Z}$.*

THEOREM 4. *For $\widehat{f} \in \mathcal{B} \cup \mathcal{B}_{2\pi}$,*

1. *the only even solutions of* (1.2) *with n even are those of* (1.5b) *with $d = 0$;*

2. *the only even solutions of* (1.2) *with n odd and for which $\widehat{f}(0) \neq 0$ are those of* (1.5b) *with $d = 0$.*

4. Even solutions of the transformed equation and their inverses. Theorems 1–4 state that the only even solutions of (1.2) with $\widehat{f}(0) \neq 0$ are those of (1.5b) with $d = 0$. We complete the proof of these theorems in this section and section 5 by deriving these solutions for the case $n = 2$ and then in section 5 by showing that these solutions also satisfy (1.2) for arbitrary n; see (5.5)–(5.9). If there are no poles of f on the real axis, then (3.14) for the Fourier transform, when $n = 2$, now reads

$$\left(k\widehat{f}(l) + l\widehat{f}(k)\right)\widehat{f}(k+l) = (k+l)\,\widehat{f}(k)\,\widehat{f}(l), \tag{4.1}$$

where for convenience we have written $k_1 = k$ and $k_2 = l$. Clearly if $\widehat{f}(k)$ is a function not identically zero, then (4.1) implies that $k/\widehat{f}(k)$ is linear so that $\widehat{f}(k)$ is constant. This, however, gives only the distributional solution where $f(x)$ is a constant multiple of $\delta(x)$. If we allow distributions, then $\widehat{f}(k) = a\delta(k)$ is also a solution of (4.1), as is $\widehat{f}(k) = a\delta(k) + b$ with the corresponding $f(x) = b\delta(x) + a$. These are the only even solutions of (4.1).

When we allow f to have a double pole at the origin, (3.14) for the Fourier transform, when $n = 2$, becomes

$$\left(k\widehat{f}(l) + l\widehat{f}(k)\right)\widehat{f}(k+l) = (k+l)\,\widehat{f}(k)\,\widehat{f}(l) + \pi^2 kl\,(k+l). \tag{4.2}$$

The Fourier transform of $1/x^2$ is $-\pi|k|$, and the identity

$$(k|l| + l|k|)\,|k+l| = (k+l)\,(|k|\,|l| + kl) \tag{4.3}$$

for all k, l ensures that the even functions $\widehat{f}(k) = \pm\pi|k|$ satisfy (4.2), with the minus sign required to satisfy the pole condition. However, $\widehat{f}(k) = \pm\pi|k|$ are not the only solutions of (4.2). We wish to consider only even functions, but also wish to include functions like $|k|$ which are not differentiable at $k = 0$. Hence we consider (4.2) defined on the subspace $k \geq 0, l \geq 0$, with all derivatives at the origin defined by one-sided derivatives. Thus in the interval $k \geq 0$, $|k|$ is defined as k with derivative 1 at the origin. Writing (4.2) as

$$\widehat{f}(l)\left(k\left(\widehat{f}(k+l) - \widehat{f}(k)\right) - l\widehat{f}(k)\right) + l\widehat{f}(k)\,\widehat{f}(k+l) = \pi^2 kl\,(k+l) \tag{4.4}$$

and then dividing by l and taking the limit as $l \to 0$ gives

$$a_0 k\widehat{f}'(k) - a_0\widehat{f}(k) + \widehat{f}^2(k) = \pi^2 k^2, \tag{4.5}$$

where $a_0 = \widehat{f}(0)$, and the solution of (4.5) with $\widehat{f}(0) = a_0$, when $a_0 \neq 0$, is

$$\widehat{f}(k) = \pi k \coth(\pi k/a_0). \tag{4.6}$$

We note here that (4.6) requires only $k^{-1}(\widehat{f}(k) - \widehat{f}(0) - k\widehat{f}'(0)) = o(1)$ as $k \to 0$. This will be the case if $\widehat{f}'(k)$ is continuous at the origin. This is provided for by property (3′a) of Definition 1, which ensures that $\widehat{f}(k)$ has a continuous second derivative. Of course for an even function with this property, $\widehat{f}'(0) \equiv 0$. The function appearing in (4.6) is the Fourier transform of the function

$$f(x) = -\frac{1}{4}a_0|a_0| / \sinh^2\left(\frac{a_0 x}{2}\right). \tag{4.7}$$

Again, apart from the addition of a constant and a scaling of x, this produces the unique (degeneration of a) $\wp$ function of imaginary period π and real period infinity.

As $a_0 \to 0$ the solution (4.6) tends to $\pm\pi|k|$, which is also the solution of (4.5) when $a_0 = 0$. This requires $k^{-1}(\widehat{f}(k) - \widehat{f}(0) - |k|\widehat{f}'(0+)) = o(1)$ as $k \to 0$, where $\widehat{f}'(0+)$ is defined to be the limit of $\widehat{f}'(k)$ as k tends to zero through positive values. This is provided for by the stronger property (3′b) of Definition 1.

For even functions $f(x)$ which are 2π-periodic in addition to having a double pole of the form $1/x^2$ at the origin following (3.18), we write $\widehat{f}(k)$ in the form

$$\widehat{f}(k) = \sum_{p=-\infty}^{\infty} a_p \delta(k-p), \quad \text{with } a_{-p} = a_p, \tag{4.8}$$

to obtain the recurrence relation for the set $\{a_K\}$,

$$(Ka_L + La_K)\,a_{K+L} = (K+L)\,a_K a_L + \pi^2 KL(K+L). \tag{4.9}$$

When $K = 0$, (4.9) is automatically satisfied with a_0 arbitrary. Writing $a_K = \pi K b_K$ for $K = 1, 2, \ldots$, we obtain

$$(b_L + b_K)\,b_{K+L} = b_K b_L + 1, \tag{4.10}$$

and with $b_1 = \frac{\beta+1}{1-\beta}$ and $L = 1$, we have

$$b_{K+1} = (b_K(\beta+1) + 1 - \beta) / (b_K(1-\beta) + \beta + 1). \tag{4.11}$$

This is easily solved to yield

$$b_K = \frac{1+\beta^K}{1-\beta^K} \qquad \text{or} \qquad a_K = \pi\,|K|\left(\frac{1+\beta^{|K|}}{1-\beta^{|K|}}\right). \tag{4.12}$$

This reproduces the result that $a_K = \widehat{f}\,(K)$ at integer values of K, where $\widehat{f}\,(K)$ is given by the continuous equation. Also the Fourier series $\sum_{-\infty}^{\infty} \pi\,|K|\,\big(\frac{1+q^{2|K|}}{1-q^{2|K|}}\big)e^{iKx}$ can be recognized as the Fourier series for the $\wp$ function again (up to the addition of a constant). Here $q = \beta^{1/2}$ is the usual notation for the nome.

At this stage we have found the solution set $\mathcal{F}_2$. The conclusion is that the only even solutions of (1.2) with $n = 2$ are those of (1.5b) with $d = 0$.

5. New solutions. We now look at the solutions of (1.2) which have $\widehat{f}\,(0) = 0$ to see if there are solutions which do not belong to the set $\mathcal{F}_2$. Lemma 1 and Theorem 2 can be easily adapted to prove that when n is odd, $\mathcal{F}_n \subseteq \mathcal{F}_3$. We first find this solution set and then prove that when n is odd, $\mathcal{F}_n = \mathcal{F}_3$. Hence we look for solutions of (3.14), with $\widehat{f}\,(0) = 0$, when $n = 3$. We wish to consider only even functions but again wish to include functions like $|k|$ which are not differentiable at $k = 0$. Hence we consider (3.14) defined on the subspace $k_q \geq 0$ $(q = 1, 2, 3)$ with all derivatives at the origin defined by one-sided derivatives. Thus in the interval $k \geq 0$, $|k|$ is defined as k with derivative 1 at the origin.

In (3.14) we write $k_1 = k$ and $k_2 = k_3 = l$ and let $l \to 0$. Then (3.14) gives

$$a_0\left(a_0 k\widehat{f}\,'(k) - a_0\widehat{f}\,(k) + \widehat{f}^{\,2}(k) - \pi^2 k^2\right) = 0, \tag{5.1}$$

where $a_0 = \widehat{f}\,(0)$. When $a_0 \neq 0$ this gives the same equation as (4.5) but is automatically satisfied when $a_0 = 0$. So in addition to the solution given in (4.6), which belongs to $\mathcal{F}_2$, we can also allow $a_0 = 0$.

When $a_0 = 0$ the next term in the expansion of (3.14) gives

$$a_1\left(\widehat{f}^{\,2}(k) - \pi^2 k^2\right) = 0, \tag{5.2}$$

where $a_1 = \widehat{f}\,'(0)$. If $a_1 \neq 0$, then (5.2) gives $\widehat{f}\,(k) = \pm\pi\,|k|$ as the only even solution. This is also the solution of (4.6) when $a_0 = 0$ and hence also belongs to $\mathcal{F}_2$. Now if we assume that $a_1 = 0$, we can write the third term in the expansion of (3.14) as

$$2\pi^2 k\widehat{f}\,'(k) + a_2\widehat{f}^{\,2}(k) - 2\pi^2\widehat{f}\,(k) = \pi^2 a_2 k^2, \tag{5.3}$$

where $a_2 = \widehat{f}\,''(0)$. The derivation of this solution requires $k^{-2}(\widehat{f}(k) - \widehat{f}(0) - k\widehat{f}\,'(0) - k^2\widehat{f}\,''(0)/2) = o(1)$ as $k \to 0$. This will be the case if $\widehat{f}\,''(k)$ is continuous at the origin and is provided for by the property (3′a) of Definition 1. Again for an even function with this property, $\widehat{f}\,'(0) \equiv 0$.

The only even solution of this equation is

$$\widehat{f}\,(k) = \pi k \tanh\left(\frac{ka_2}{2\pi}\right), \tag{5.4}$$

which automatically has $f''(0) = a_2$. This of course is a necessary requirement, and we need to check that this is a solution of (3.14).

We rewrite (3.14) as

(5.5)
$$\widehat{S}_n \equiv \left(\sum_{j=1}^{n} \frac{k_j}{\widehat{f}\,(k_j)+\pi k_j} \prod_{q=j+1}^{n} \left(\frac{\widehat{f}\,(k_q)-\pi k_q}{\widehat{f}\,(k_q)+\pi k_q}\right)\right) \frac{\left(\widehat{f}\left(\sum\limits_{j=1}^{n} k_j\right)+\pi \sum\limits_{j=1}^{n} k_j\right)}{\sum\limits_{j=1}^{n} k_j} - 1 = 0.$$

Substituting $\widehat{f}\,(k) = \pi k \tanh(ak)$ into $\widehat{S}_n$ gives

$$\widehat{S}_n = -\left(\sum_{j=1}^{n} \left(e^{2ak_j}+1\right) \prod_{q=j}^{n} \left(-e^{-2ak_q}\right)\right) \frac{\exp\left(2a \sum\limits_{j=1}^{n} k_j\right)}{\exp\left(2a \sum\limits_{j=1}^{n} k_j\right)+1} - 1. \tag{5.6}$$

The first part of this expression can be written as

$$\widehat{S}_n^{(1)} = \sum_{j=1}^{n} a_{j+1} - a_j \equiv a_{n+1} - a_1, \tag{5.7}$$

where $a_j = \prod_{q=j}^{n} \left(-e^{-2ak_q}\right)$, so that $\widehat{S}_n^{(1)} = (-1)^{n+1} \exp\left(-2a \sum_{j=1}^{n} k_q\right) + 1$. Hence

$$\widehat{S}_n = \frac{\left((-1)^{n+1}-1\right)}{\exp\left(2a \sum\limits_{j=1}^{n} k_j\right)+1}. \tag{5.8}$$

This immediately gives $\widehat{S}_n \equiv 0$ whenever n is odd. Hence $\widehat{f}\,(k) = \pi k \tanh(ak)$, with a arbitrary, is a solution for all equations $S_n = 0$ when n is odd. It is also evident from (5.8) that this solution does not satisfy $S_n = 0$ for n even.

We also note that substituting $\widehat{f}\,(k) = \pi k \coth(ak)$ into (5.5) changes (5.6) to

$$\widehat{S}_n = \left(\sum_{j=1}^{n} \left(e^{2ak_j}-1\right) \prod_{q=j}^{n} e^{-2ak_q}\right) \frac{\exp \sum\limits_{j=1}^{n} (2ak_j)}{\exp\left(\sum\limits_{j=1}^{n} 2ak_j\right)-1} - 1. \tag{5.9}$$

The change in signs now means that $\widehat{S}_n \equiv 0$ for all n, showing that $\widehat{f}_1\,(k) = \pi k \coth(ak)$, with a arbitrary, satisfies (3.14) for all n, as claimed earlier.

If we write $a = \frac{1}{2}\pi/\alpha$ so that $\widehat{f}\,(k) = \pi k \tanh\left(\frac{1}{2}\pi k/\alpha\right)$, then this is the Fourier transform of the function $f(z) = -\alpha|\alpha|\cosh \alpha z/\sinh^2 \alpha z$, so α must be negative to satisfy the pole condition that $z^2 f(z) \to 1$ as $z \to 0$. However, since the coefficient of the double pole is in fact arbitrary, we have $f(z) = \beta \cosh \alpha z/\sinh^2 \alpha z$ satisfying (1.2) for all odd values of n. (We will give an alternate way of discovering this solution in section 7.)

The even periodic solutions, which satisfy the modification of (3.14) when $f(z)$ has an array of double poles of the form $(z-2p\pi)^{-2}$ at the points $z = p\pi$, $p = 0$,

$\pm 1, \pm 2, \ldots$, can be written as

$$\widehat{f}(k) = \sum_{p=-\infty}^{\infty} a_p \delta(k-p), \quad \text{with } a_{-p} = a_p, \tag{5.10}$$

as in (3.18). Again, (3.14) is now to be satisfied at all integer values of $\{k_q\}$ with $\widehat{f}(k_q)$ replaced by a_{k_q}. Hence

$$\widehat{S}_n \equiv \sum_{j=1}^{n} \left[K_j \prod_{q=1}^{j-1} \left(a_{K_q} + \pi K_q\right) \prod_{q=j+1}^{n} \left(a_{K_q} - \pi K_q\right) \right] \left(a_{\sum_{q=1}^{n} K_q} + \pi \sum_{q=1}^{n} K_q \right) - \left(\sum_{j=1}^{n} K_j \right) \left(\prod_{q=1}^{n} \left(a_{K_q} + \pi K_q\right) \right) = 0. \tag{5.11}$$

To solve (5.11) to obtain the solutions for $\widehat{S}_3 = 0$, we proceed as in section 4. If a_0 is not equal to zero, we can put $K_3 = 0$ and recover the solutions (4.12). However, if $a_0 = 0$, then $\widehat{S}_3 \equiv 0$ for all K_1 and K_2 if $K_3 = 0$. Writing down all the equations for $K_j \geq 1$ we find that if $a_1 = \pi(\beta+1)/(1-\beta)$, the odd terms are given by

$$a_{2K+1} = \pi(2K+1) \frac{\left(1+\beta^{2K+1}\right)}{1-\beta^{2K+1}}, \qquad K \geq 0. \tag{5.12}$$

This is established in the same way that (4.12) is obtained from (4.9). Writing $a_K = \pi K b_K$ for $K = 1, 2, \ldots$, and with $K_1 = K_2 = 1$ and $K_3 = 2K+1$ we obtain

$$b_{2K+1} = \left(b_{2K-1}\left(\beta^2+1\right) + 1 - \beta^2\right) / \left(b_{2K-1}\left(1-\beta^2\right) + \beta^2 + 1\right), \tag{5.13}$$

which is solved to get (5.12). However, the even terms depend on the choice of a_2, which must take one of the values

$$a_2 = 2\pi \frac{\left(1+\beta^2\right)}{1-\beta^2} \tag{5.14a}$$

or

$$2\pi \frac{\left(1-\beta^2\right)}{1+\beta^2}. \tag{5.14b}$$

This is obtained by choosing $K_1 = 1, K_2 = 2$, and $K_3 = 2$, with a_1 and a_5 given from (5.12). This yields a quadratic for a_2 with (5.14) as solution. The first choice in (5.14) gives

$$a_{2K} = 2\pi K \frac{\left(1+\beta^{2K}\right)}{1-\beta^{2K}}, \qquad K \geq 0, \tag{5.15}$$

and the second

$$a_{2K} = 2\pi K \frac{\left(1-\beta^{2K}\right)}{1+\beta^{2K}}, \qquad K \geq 0. \tag{5.16}$$

Again these results are obtained from the choice of $K_1 = K_2 = 1$ and $K_3 = 2K$, which, using the definition of b_K, gives

$$b_{2K+2} = (2b_1 + b_{2K} + b_1^2 b_{2K})/(2b_1 b_{2K} + 1 + b_1^2), \tag{5.17}$$

which is solved to give (5.15) or (5.16), depending on the choice of a_2.

The results (5.14a) and (5.15) are equivalent to (4.12), while writing $\beta = -\tilde{\beta}$ shows that (5.14b) and (5.16) are equivalent to

$$a_K = \pi |K| \left(\frac{1 - \beta^{|K|}}{1 + \beta^{|K|}} \right) \qquad \text{for all } K. \tag{5.18}$$

This reproduces the result of (5.4) at integer values of k. The proof that $\{a_K\}$ satisfies $\widehat{S}_n$ for all n is identical to the proof in the continuous case (see (5.5)–(5.8)).

The corresponding inverse $f(z)$, obtained from $\widehat{f}(k)$, can be constructed either by taking the Fourier inverse of $\widehat{f}(k)$ or by the infinite sum defined by (3.21) using the function $\beta \cosh \alpha z / \sinh^2 \alpha z$. This function must be one of $\mathrm{cn}/\mathrm{sn}^2$, $\mathrm{dn}/\mathrm{sn}^2$, or $\mathrm{cndn}/\mathrm{sn}^2$. There are two reasons why there are three functions representing the solution set. The first is that if all the parameters defining the elliptic function are real, then the transformation $z \to iz$ permutes these three functions according to Jacobi's imaginary transformation $\mathrm{sn}(iz, k) \to i\ \mathrm{sn}\left(z, 1-k^2\right)/\mathrm{cn}\left(z, 1-k^2\right)$, $\mathrm{cn}(iz, k) \to 1/\mathrm{cn}\left(z, 1-k^2\right)$, and $\mathrm{dn}(iz, k) \to \mathrm{dn}\left(z, 1-k^2\right)/\mathrm{cn}\left(z, 1-k^2\right)$. Second, Jacobi's real transformation defines the elliptic function for the parameter $k < 0$ or $k > 1$ in terms of elliptic functions with a scaled independent variable and parameter k in the range $0 < k < 1$. Again the effect is to permute the three functions.

The conclusion is that the even solutions of (1.2) with n odd fall into two categories. One is the set defined by (1.5b) with $d = 0$, which satisfy (1.8) subject to $z^2 f(z) \to$ a constant as $z \to 0$. There is also a further set which may be expressed in terms of the $\wp$ function and also in terms of the Jacobian elliptic functions as given in Theorem 1. We have now proven Theorem 1.

6. Noneven solutions. At this stage we have established the theorem. For completeness we show how the general solutions of (1.1) may be obtained using the Fourier transform method. The even solutions have already been obtained, and now we consider the noneven solutions. We now assume that the function f in (1.4) is not necessarily even but is bounded on the real axis, but otherwise satisfies the conditions given in definition 1. Then the method of section 3 can be adapted to give the equation

$$k\widehat{f}(-l)\,\widehat{f}(-k-l) + l\widehat{f}(k)\,\widehat{f}(k+l) - (k+l)\,\widehat{f}(-k)\,\widehat{f}(l) = 0. \tag{6.1}$$

Equation (6.1) is a rather complicated functional equation when $\widehat{f}(-k) \neq \widehat{f}(k)$. A Taylor series method produces a three-parameter family of solutions. Two parameters are as a consequence of the fact that if $\hat{g}(k)$ is a solution, so is $\widehat{f}(k) = a\widehat{g}(bk)$ for all constants a and b. This is as a result of the scaling symmetries of the original equation (6.1). So essentially there is a one parameter family of solutions. However, it is not easy to recognize the solution from its series. The method which appears to give the most simple solution is the following. We decompose $\widehat{f}$ into an even and odd function of the form

$$\widehat{f}(k) = f_1(k) + tk f_2(k), \tag{6.2}$$

where f_1 and f_2 are even functions of k and t can be ± 1. Substituting this expression into (1.4) yields an equation of the form

$$A(k,l) + tB(k,l) + t^2 C(k,l) = 0, \tag{6.3}$$

and since t can be either ± 1 this gives two equations

$$A + C = 0 \qquad \text{and} \qquad B = 0. \tag{6.4}$$

Now we assume that l is small and expand (6.4) as a power series in l. Equating the coefficients of the powers of l to zero gives a series of differential equations involving $f_1(k)$ and $f_2(k)$. From the first two equations we obtain

$$2(f_2(k) - a_2) f_1(k) = k\, a_1\, f_2'(k) \tag{6.5}$$

and

$$ka_1 f_1'(k) + f_1^2(k) + k^2 f_2^2(k) - f_1(k)\, a_1 + 2k^2 a_2 f_2(k) = 0, \tag{6.6}$$

where

$$a_1 = f_1(0) \qquad \text{and} \qquad a_2 = f_2(0). \tag{6.7}$$

Eliminating $f_1(k)$ we obtain

$$a_1^2 \left(2(f_2 - a_2) f_2'' - f_2'^2\right) + 4f_2\left(f_2^3 - 3a_2^2 f_2 - 2a_2^3\right) = 0, \tag{6.8}$$

and one integration yields

$$3a_1^2 f_2'^2 = 4(a_2 - f_2)\left(ca_2^3 + f_2^3 + 3a_2 f_2^2\right), \tag{6.9}$$

where c is an arbitrary constant. This constant is the third parameter referred to above. Examination of the cubic in (6.9) shows that for all (real) constants c other than $c = 0$ or -4 we have an oscillatory solution for f_2. These solutions must be rejected on the grounds that if the original function $f(x)$ is bounded, $\widehat{f}(k)$ must tend to zero as $k \to \pm\infty$. When $c = -4$, $f_2 \equiv a_2$, and (6.5) is automatically satisfied, (6.6) then yields

$$f_1(k) = \sqrt{3} a_2 k \cot\left(\sqrt{3} a_2 k / a_1\right). \tag{6.10}$$

Again this does not represent the Fourier transform of a bounded function $f(x)$. However, it does illustrate the general feature of all periodic solutions of (6.9). When $c \neq 0$ the solution of (6.9) with $f_2(0) = a_2$ is an elliptic function which is even. Clearly, when $c \neq -4$, $f_2(k) - a_2$ has zeros when $k = 0$ or an integer multiple of the period of f_2. Since (6.9) implies that $f_2' = 0$ and $f_2'' \neq 0$ when $f_2 = a_2$, at each zero other than $k = 0$, $f_1(k)$ has a simple pole. Hence f_1 has a periodic array of poles, apart from $k = 0$, where the singularity is removable. This feature is illustrated by the function appearing in (6.10).

When $c = 0$, we can solve (6.9) with $f_2(0) = a_2$ to get

$$f_2(k) = \frac{3a_2}{2\cosh 2\alpha k + 1}, \tag{6.11}$$

where $\alpha = a_2/a_1$, so that

$$f_1(k) = \frac{3a_2 k \cosh \alpha k}{\sinh \alpha k \,(2\cosh \alpha k + 1)}. \tag{6.12}$$

The two possibilities for $\widehat{f}\,(k)$ are then

$$\widehat{f}\,(k) = f_1\,(k) \pm k f_2\,(k)\,, \tag{6.13}$$

giving

$$\widehat{f_1}\,(k) = \frac{6a_2 k e^{4\alpha k}}{e^{6\alpha k} - 1} \qquad \text{and} \qquad \widehat{f_2}\,(k) = \frac{6a_2 k e^{2\alpha k}}{e^{6\alpha k} - 1}\,. \tag{6.14}$$

If we normalize by choosing $6a_2 = \pi a_1$ and $a_1 = -2$, then $\widehat{f_1}$ is the Fourier transform of the function $1/\sinh^2(x - i\pi/3)$ and $\widehat{f_2}$ is the Fourier transform of $1/\sinh^2(x + i\pi/3)$. Apart from the addition of a constant these are $\wp$ functions, with periods (∞, π), and a double pole at the origin.

Thus, apart from the double pole at the origin and the shift by one-third and two-thirds of the imaginary period of the $\wp$ function, there is a unique solution of (1.4) which is bounded on $(-\infty, \infty)$ and tends to zero at ∞. An interesting limit is $a_2 \to 0$, which gives $f_2 \equiv 0$ and $f_1 = a_1$. This is also seen to be the only solution of (6.5) and (6.6) when $a_2 = 0$, subject to the condition that $f_1(0) = a_1$, $f_2\,(0) = a_2 = 0$, and $f_2 \to 0$ as $k = \infty$. The inverse of the Fourier transform $\widehat{f}\,(k) \equiv a_1$ is no longer a function but the distribution $a_1\delta\,(x)$ which can be viewed as the limit

$$a_1\delta\,(x) = \lim_{\alpha \to 0} -\frac{a_2}{\alpha \sinh^2(-x/\alpha + \pi i/3)},$$

where $\pi a_1 \alpha = 6a_2$.

The only other regular solutions of (1.4), on the real axis, are ones that are either not bounded at ∞ or are oscillatory. Such functions have either distributional Fourier transforms or are such that the no conventional Fourier transform exists. For example, the Fourier transform of $e^{i\alpha x}$ is $2\pi\delta\,(k - \alpha)$ and as a distribution $f = 2\pi\delta\,(k - \alpha)$ satisfies (6.1), since

$$\begin{aligned} 4\pi^2 k\delta\,(-l-\alpha)\ \delta\,(-k-l-\alpha) + 4\pi^2 l\delta\,(k-\alpha)\,\delta\,(k+l-\alpha) \\ - 4\pi^2\,(k+l)\,\delta\,(-k-\alpha)\,\delta\,(l-\alpha) \equiv 0. \end{aligned} \tag{6.15}$$

This is because $k\ \delta\,(-l-\alpha)\ \delta\,(-k-l-\alpha)$ is zero unless $l + \alpha = l + k + \alpha = 0$ or $k = 0,\ l = -\alpha$, when the coefficient of the product of δ functions is zero. A more formal proof can be obtained by defining the inner products $\langle u, v\rangle_{k,l}$ and $\langle u, v\rangle_k$ by

$$\langle u, v\rangle_{k,l} \equiv \iint_{\mathbb{R}^2} u\ v\ dkdl \qquad \text{and} \qquad \langle u, v\rangle_k \equiv \int_{\mathbb{R}} u\ v\ dk, \tag{6.16}$$

with a similar definition for $\langle u, v\rangle_l$. Then

$$\begin{aligned} \langle k\delta\,(-l-\alpha)\,\delta\,(-k-l-\alpha)\,,\ F\,(k,l)\rangle_{k,l} &= \langle \delta\,(-l-\alpha)\,\delta\,(-k-l-\alpha)\,,\ \ kF\,(k,l)\rangle_{k,l} \\ &= \langle \delta\,(-l-\alpha)\,,\ \ -(l+\alpha)\ F\,(-l-\alpha,l)\rangle_l \\ &= 0. \end{aligned} \tag{6.17}$$

Although this only shows that $e^{i\alpha x}$ is a solution of (1.4) by analogy, we can also have $e^{\alpha x}$, although this does not have a Fourier transform, except formally by analytic continuation.

The Fourier transform of x is $2\pi i\delta'(k)$, and using (6.16) it is straightforward to prove that $k\delta'(-l)\,\delta'(-k-l)+l\delta'(k)\,\delta'(k+l)$ and $(k+l)\,\delta'(-k)\,\delta'(l)$ are identical distributions. For example

$$\begin{aligned}\langle (k+l)\,\delta'(-k)\,\delta'(l)\,,\ \ F(k,l)\rangle_{k,l} &= \langle \delta'(-k)\,\delta'(l)\,,(k+l)\,F(k,l)\rangle_{k,l}\\ &= \langle \delta'(-k)\,,-F(k,0)-kF_l(k,0)\rangle_k\\ &= -F_k(0,0)-F_l(0,0)\,.\end{aligned}\tag{6.18}$$

A similar calculation shows that

$$\langle k\delta'(-l)\,\delta'(-k-l)\,,\ F(k,l)\rangle_{k,l}+\langle l\delta'(k)\,\delta'(k+l)\,,\ F(k,l)\rangle_{k,l}\ =-F_k(0,0)-F_l(0,0)\,.\tag{6.19}$$

In addition, we can also see that if $\widehat{f}(k)$ is not continuous, $\widehat{f}(0)$ can be arbitrary since (6.1) is satisfied automatically when either k, l, or $k+l$ equals zero. In particular, we can add an arbitrary multiple of $\delta(k)$ to any solution of (6.1). This corresponds to adding a constant to any solution of (1.4).

The above distributional solutions cover the solutions of (1.5a). To cater to the noneven periodic solutions of (1.5b) that are also bounded on the real axis, we assume that the solutions $f(x)$ are 2π-periodic, and following (3.17) we write

$$f(x)=\frac{1}{2\pi}\sum_{p=-\infty}^{\infty}a_p e^{ipx}\tag{6.20}$$

to obtain

$$Ka_{-L}a_{-K-L}+La_Ka_{K+L}-(K+L)\,a_{-K}a_L=0\tag{6.21}$$

for all integer values of (K,L). Since $\widehat{f_1}(k)$ and $\widehat{f_2}(k)$ defined by (6.13) satisfy the continuous version of (6.21), it is clear that the solution $a_K=\widehat{f_1}(K)$, $K\neq 0$ with a_0 arbitrary, will satisfy (6.21). This solution for a_K can also be obtained directly from (6.21) by successively solving all equations with $1\le|K|\le N$, $1\le|L|\le N$ for $N=1,2,3,\ldots$, the equations where $K=0$ or $L=0$ being automatically satisfied. Apart from the fact that a_0 is undetermined and is thus arbitrary, this process, as in the continuous case, produces a three-parameter family of solutions with a_{-2},a_{-1}, and a_1 arbitrary. If we choose $a_{-1}=\beta\alpha^2/\left(\alpha^6-1\right)$ and $a_1=\beta\alpha^4/\left(\alpha^6-1\right)$, then the choice $a_{-2}=\beta\alpha^4/\left(\alpha^{12}-1\right)$ yields $a_K=\beta K\alpha^{4K}/\left(\alpha^{6K}-1\right)$, which is clearly equivalent to $\widehat{f_1}(K)$ when $\beta=6a_2$ and $\alpha=\exp(a_2/a_1)$. Since all continuous solutions $\widehat{f}(k)$ give a corresponding solution $a_K=\widehat{f}(K)$, via (3.22), we presume that other choices of a_{-2} give solutions for a_K which are oscillatory and do not tend to zero as $K\to\infty$. As in the continuous case this gives Fourier series which do not come from a continuous function of x.

It is also easy to show that there are additional solutions of (6.21). If the set $S_1=\{a_K\}$ solves (6.21), then so does the set $S=\left\{\gamma^K a_K\right\}$, provided $\gamma^{3K}=1$ for all integer values of K. This gives two additional solutions $\left\{e^{2\pi Ki/3}a_K\right\}$ and $\left\{e^{4\pi Ki/3}a_K\right\}$. If $f_1(x)$ is the 2π-periodic function whose Fourier series is given by the set S_1, then

these two solutions are from the functions $f_1\left(x+\frac{2}{3}\pi\right)$ and $f_1\left(x+\frac{4}{3}\pi\right)$, which are the original function shifted by one-third and two-thirds of its period, respectively.

In this section we have shown that by requiring f to be bounded on any compact subset of the real axis, we can recover all the solutions of (1.5a,b) that have this boundedness property by taking the appropriate Fourier transform of (1.4). We have also shown that there are no other solutions f which have this boundedness property. However, by taking the appropriate limit we have also shown that the δ function is also a solution and have recovered this from the transformed equation. We have not obtained the $\wp$ function solutions, which have double poles on the real axis. This is because the transforms of these functions do not satisfy the transformed equation (6.1). These follow from an analysis similar to that given in section 4.

7. Series solution approaches. We conclude this paper by describing two methods based on a series approximation for studying (1.2). Although they are incomplete, and this was the reason for developing the new methods already described, it is helpful to see how our new solutions arise in this setting. In doing this, we see that the $n = 2$ and 3 solutions can be derived assuming just the meromorphic nature, pole assumptions, and evenness of the solutions contained in the first two items of Definitions 1 and 2.

Method 1: Obtaining a series solution. One method of attempting to prove the conjecture is to assume all the $x_i, i = 1, \ldots, n+1$, are small and of the same order, so that we write $x_i = t\zeta_i$. Then we assume that all the even functions, such as $f(t\zeta)$, can be expressed as a power series in t with ζ as an order-one parameter in the form

$$f(t\zeta) = \sum_{j=0}^{\infty} a_j c_{2j-2} (t\zeta)^{2j-2}. \tag{7.1}$$

The constants c_j are given by $c_j = 1$ if $j < 0$ and $c_j = 1/j!$ if $j \geq 0$ and are included for convenience. The series (7.1) allows for a double pole at the origin, which can easily be shown to be the only allowable singularity. The coefficients a_j are determined by equating to zero the coefficients of the powers of t^{2j} in the subsequent expansion of (1.2). These coefficients are of course functions of ζ_i as well as a_j. However, each coefficient factorizes into a product of homogeneous polynomials in ζ_i, independent of a_j and a factor dependent on the a_j only. Equating this coefficient to zero successively determines a_j for all $j \geq 4$ in terms of a_0, a_1, a_2, and a_3, which are arbitrary. This process can be completed to any desired order, J, if the expansion (7.1) is truncated at a suitable finite value. Substitution of this finite polynomial into (1.7) shows, for the cases where the method works, that for a suitable choice of $\{A,\ B,\ C,\ D\}$, (1.7) can be satisfied to any desired order.

The proof of the above statement is, of course, incomplete: the form of the general term, a_j, is not obtained, and hence we cannot show that the full expansion (7.1) satisfies (1.7). This method works for $n = 2$ and $n = 4$, but for values of $n \geq 5$ the amount of algebra involved becomes so large that even a Maple calculation cannot handle the details. When $n = 3$, this method does not work completely in that it leaves a_4 arbitrary. Our earlier analysis has shown that this is not the case.

It is interesting to note that the same procedure works for solutions of (1.1),

although the assumption of evenness is not required. Hence we look for a solution

$$f(t\zeta) = \sum_{j=0}^{\infty} a_j c_{j-2} (t\zeta)^{j-2}. \tag{7.2}$$

If we take $a_0 \neq 0$, the process automatically gives $a_{2j+1} \equiv 0$, producing the same even function as obtained via (7.1) for the solutions of (1.2). However, if we take $a_0 = 0$, then we find that $a_1 \equiv 0$, and (7.2) is then a Taylor series. We proceed as above to produce the coefficients a_j for all $j \geq 7$ in terms of a_2 to a_6 and show that (1.7) can be satisfied to any order. However, a_5 is not arbitrary and is given by

$$a_3 a_5 = a_4^2. \tag{7.3}$$

This condition is automatically satisfied for all functions in the set (1.5a). However (7.3) is also equivalent to the condition $\wp''^2(d) = 12\wp(d)\wp'^2(d)$, which is satisfied when d is one-third or two-thirds of any period of $\wp(z)$. This gives the required condition on the constant d appearing in (1.5b) when f belongs to this solution set.

Method 2: Obtaining a series of differential equations. An alternative method is to assume that one variable, for example x_1, is not small and write $x_1 = x$ and $x_i = t\zeta_i$, $2 \leq i \leq n+1$, together with (7.1), although x_1 can be replaced by any linear combination of the x_i, as long as n variables are scaled by t. The expansion of $f(x - t\zeta)$ then naturally produces coefficients of t^j, which are functions of f and its higher derivatives. The coefficients of t^j, in the expansion of (1.2) when equated to zero, now yield differential equations which must be satisfied by $f(x)$ but contain the "arbitrary" constants a_j.

For the case $n = 2$ the first equation is

$$a_0 f''' + 12 a_1 f' - 12 f f' = 0. \tag{7.4}$$

This proves to be sufficient to determine a differential equation for f in the sense that the elimination of the constants a_0 and a_1 by differentiation gives (1.8). The higher order coefficients of t^j produce equations similar to (7.4) but contain the constants a_j with $j \geq 4$, which are not arbitrary. In this case eliminating all the constants by differentiation will yield an equation which is still necessary but not sufficient to determine f. If we seek a series solution to (7.4) by looking for a series solution of the form $f(x) = \sum_{j=0}^{\infty} b_j c_{2j-2} x^{2j-2}$, then we obtain $b_0 = a_0$ and $b_1 = a_1$, with b_2 and b_3 arbitrary. The coefficients $b_j, j \geq 4$, are then determined by these four constants, the recurrence relation being the same as in Method 1. The equations which contain a_2 and a_3 also require $b_2 = a_2$ and $b_3 = a_3$ and so also reproduce the series (7.1).

This method is more complete than the previous method in that it does yield a necessary differential equation, namely, (1.8), that all even solutions of (1.2) for $n = 2$ must satisfy. However, we cannot prove that all the differential equations produced by the expansion are satisfied. But this is not a problem since it is easy to verify, by substitution, that all even solutions of (1.5b) satisfy (1.2) for this case.

Again, the method can be adapted to obtain the solution of (1.1), which are not even, using (7.2) with $a_0 = a_1 = 0$. The procedure is the same except that because $f(\zeta)$ is not even, we obtain differential equations containing $f(\zeta)$ and $f(-\zeta)$. When one of these functions is eliminated we can then show that either (1.6) or (1.8) is satisfied. However, the differentiations required to eliminate the constants a_j mean that the necessary condition (7.3) is not recovered.

For the case $n = 3$ the situation is more complicated since the series of equations similar to (7.4) only produce necessary equations for f. In order to show that f satisfies (1.7) we need to take two equations similar to (7.4) and find the set of solutions common to both differential equations. The algebraic details are quite complicated and require a Maple calculation. The details are not repeated here, but a summary of the results is given. A copy of the Maple program which produces these results with further explanation can be obtained by contacting the first author.

The two equations, whose common solution satisfies (1.2) for $n = 3$ are given by

$$120\,a_2\,f'\,f + 5\,a_0\,f'''\,f'' + 60\,f''\,f'\,a_1 - a_0\,f^{(5)}\,f = 0 \tag{7.5}$$

and

$$\begin{aligned} 0 = {} & 504\,a_3\,f'\,f^2 + 1080\,f'\,a_2\,f''\,f + 36\,f^{(iv)}\,f'\,a_1\,f + 15\,f'\,a_0\,f'''^2 + 180\,f'\,f''^2\,a_1 \\ & + 180\,f'''\,f'^2\,a_1 - 3\,a_0\,f^{(v)}\,f'^2 + 360\,f'^3\,a_2 + 15\,f'\,a_0\,f^{(iv)}\,f'' - 60\,f''\,a_1\,f'''\,f \\ & - 12\,a_0\,f^{(iv)}\,f'''\,f + 240\,f'''\,a_2\,f^2 + +a_0\,f^{(v)}\,f''\,f. \end{aligned} \tag{7.6}$$

To determine the equation or equations for these common solutions, we effectively eliminate the arbitrary constants a_0, a_1, a_2, and a_3. These constants are the ones that appear in the expansion (7.1) used to derive (7.5) and (7.6).

Before we consider the common solution to these equations, we investigate the form of these solutions by looking for a common solution of the form (7.1) but with the constants a_j replaced by b_j. The result again shows that $b_0 = a_0, b_1 = a_1, b_2 = a_2$, and $b_3 = a_3$. Initially b_4 is not determined, but b_5 and b_6 are determined in terms of $b_j, j \le 4$. However, we get two different values of b_7, and equating these values gives an equation which has two solutions, namely,

$$b_3 = \frac{60\left(2b_1^2 - b_0 b_2\right)}{7b_0^2} \tag{7.7a}$$

or

$$b_4 = \frac{60b_2^2}{b_0}. \tag{7.7b}$$

If we choose (7.7b), then in turn we find both equations require a common value for b_j for $j \ge 7$, and the series generates, as in the previous method, a solution of (1.7) or (1.8) and so belongs to the solution set (1.5).

However, if we choose (7.7a) to be satisfied, then the two equations have different solutions for b_8, and equating these values determines b_4. This value differs from that given in (7.5b) and hence generates a solution $f(z) = f_1(z)$, with three arbitrary constants, which is not a solution of (1.7) or (1.8), unless

$$b_2 = \frac{6b_1^2}{5b_0}\,. \tag{7.8}$$

It is easy to verify that when (7.8) is satisfied, any finite truncation of the series is equal to the same truncation of the solution

$$f_2(z) = b_1 \wp\left(\sqrt{\frac{b_1}{b_0}}z, 12, 8\right) + b_1. \tag{7.9}$$

This is a subset of the general solution of (1.8), which may be expressed as

$$f_2(z) = 3b_1/\sin^2\left(\sqrt{\frac{3b_1}{b_0}}z\right) \qquad \text{or} \qquad f_2(z) = -3b_1/\sinh^2\left(\sqrt{-\frac{3b_1}{b_0}}z\right),$$

depending on whether $\frac{3b_1}{b_0}$ is greater than, or less than, zero, respectively. Since the choice of (7.7a) requires a specific choice of b_3 and hence a_3, it may be thought that this solution fails to be a solution of the full equation (1.2). This supposition proves, however, to be erroneous.

We can also determine the equation for the common solution of (7.5) and (7.6) by effectively eliminating the arbitrary constants a_0, a_1, a_2, and a_3. The details are contained in the Maple program referred to earlier and are not given here. Eliminating these constants results in an equation which factorizes into a number of terms, similar to the factorization encountered in [2]. One finds that f must satisfy one of the equations

$$f' = 0 \qquad \text{or} \qquad f = \text{constant}, \tag{7.10}$$

clearly a solution, but not one of interest, and

$$f'''f - f'f'' = 0 \qquad \text{or} \qquad f''/f = \text{constant}. \tag{7.11}$$

We find that this equation satisfies (7.5) and (7.6) only if the constant appearing in (7.11) is zero. Thus the only even solution again is $f =$ constant. A further factor may be recognized as (1.8) with solution (1.5b), with $d = 0$ for an even solution.

The final factor is found to be a fifth order nonlinear, ordinary differential equation which is cubic in $f^{(v)}$ and contains over one hundred terms. It is easily verified that the series solution $f = f_1(z)$ defined earlier by taking the choice of (7.7a) for b_3 satisfies this equation. While at first sight it would seem unlikely that we could obtain the most general even solution which has a double pole at $z = 0$ of such an equation, we have shown in section 5 that the Fourier transform of the nonperiodic solution is simple and easily inverted to give

$$f(z) = a\cosh bz/\sinh^2 bz \tag{7.12}$$

for arbitrary constants a and b. This coincides with the series for $f_1(z)$ upon choosing $a = 6b_1$ and $b = \sqrt{6b_1/b_0}$. (This further requires the choice $b_2 = -7ab^2/60 \equiv -21b_1^2/5b_0$.) With this knowledge, one may guess that the periodic solutions are of the form $a\ \mathrm{cn}(bz,k)\,\mathrm{sn}^2(bz,k)$, $a\ \mathrm{dn}(bz,k)/\mathrm{sn}^2(bz,k)$, or $a\ \mathrm{cn}(bz,k)\,\mathrm{dn}(bz,k)/\mathrm{sn}^2(bz,k)$, or the alternate forms given in Theorem 1. These are all periodic equivalents of (7.12), where cn, dn, and sn are the Jacobian elliptic functions and k is the modulus. It is then possible to verify by substitution that the most general solution of the fifth order equation of the form required may be expressed as

$$f_1(z) = \sqrt{b_0^2 W^2(z) + 2b_0b_1W(z) - b_1^2 + \frac{5}{3}b_0b_2}, \tag{7.13}$$

where

$$W(z) = \wp\left(z, -\frac{20\left(b_0b_2 - 3b_1^2\right)}{3b_0^2}, \frac{8b_1\left(5b_0b_2 - 3b_1^2\right)}{3b_0^3}\right). \tag{7.14}$$

The quadratic in (7.13) has a double root when $6b_1^2 = 5b_0b_2$, which corresponds to the condition (7.8), which gives the function $f_2(z)$, defined by (7.9). Thus $f_2(z)$ is the two-parameter family of common solutions to (1.8) and the fifth order equation coming from elimination, which includes the solution b_0z^{-2} in the limit $b_1 \to 0$.

If the equation for the $\wp$ function is written as

$$\wp'^2 = 4\wp^3 - g_2\wp - g_3 \equiv 4(\wp - e_1)(\wp - e_2)(\wp - e_3) \tag{7.15}$$

in the usual notation, then the quadratic in (7.13) divides the cubic on the right-hand side of (7.15). When e_1 and e_3 are the common roots, (7.13) can be simplified to give the solution $f_1(z) = a\ \mathrm{cn}(bz,k)/\mathrm{sn}^2(bz,k)$ for suitable choices of a, b, and k in terms of b_0, b_1, and b_2. When e_2 and e_3 are the common roots, we recover the solution $f_1(z) = a\ \mathrm{dn}(bz,k)/\mathrm{sn}^2(bz,k)$, while the third choice (e_1, e_2) gives $f_1 = a\ \mathrm{cn}(bz,k)\,\mathrm{dn}(bz,k)/\mathrm{sn}^2(bz,k)$. These are our new solutions.

Acknowledgments. We wish to thank A. M. Davie for helpful discussions and for sharing his insight. One of the authors (H. W. B.) wishes to thank the Newton Institute for support during the completion of this work.

REFERENCES

[1] H. W. Braden, *Rigidity, functional equations and the Calogero-Moser model*, J. Phys., A34 (2001), pp. 2197–2204.

[2] H. W. Braden and J. G. B. Byatt-Smith, *On a functional differential equation of determinantal type*, Bull. London Math. Soc., 31 (1999), pp. 463–470.

[3] H. W. Braden and V. M. Buchstaber, *Integrable systems with pairwise interactions and functional equations*, Rev. Math. Math. Phys., 10 (1997), pp. 121–166.

[4] H. W. Braden and V. M. Buchstaber, *The general analytic solution of a functional equation of addition type*, SIAM J. Math. Anal., 28 (1997), pp. 903–923.

[5] H. W. Braden and I. M. Krichever, eds., *Integrability: The Seiberg-Witten and Whitham Equations*, Gordon and Breach, Amsterdam, 2000.

[6] M. Bruschi and F. Calogero, *General analytic solution of certain functional equations of addition type*, SIAM J. Math. Anal., 21 (1990), pp. 1019–1030.

[7] V. M. Buchstaber and A. M. Perelomov, *On the functional equation related to the quantum three-body problem*, in Contemporary Mathematical Physics, Amer. Math. Soc. Transl. Ser. 2, 175, AMS, Providence, RI, 1996, pp. 15–34.

[8] V. M. Buchstaber and I. M. Krichever, *Vector addition theorems and Baker-Akhiezer functions*, Teoret. Mat. Fiz., 94 (1993), pp. 200–212.

[9] F. Calogero, *On a functional equation connected with integrable many-body problems*, Lett. Nuovo Cimento, 16 (1976), pp. 77–80.

[10] F. Calogero, *One-dimensional many-body problems with pair interactions whose exact ground-state is of product type*, Lett. Nuovo Cimento, 13 (1975), pp. 507–511.

[11] B. A. Dubrovin, A. S. Fokas, and P. M. Santini, *Integrable functional equations and algebraic geometry*, Duke Math. J., 76 (1994), pp. 645–668.

[12] E. Gutkin, *Integrable many-body problems and functional equations*, J. Math. Anal. Appl., 133 (1988), pp. 122–134.

[13] S. N. M. Ruijsenaars and H. Schneider, *A new class of integrable systems and its relation to solitons*, Ann. Physics, 170 (1986), pp. 370–405.

[14] B. Sutherland, *Exact ground-state wave function for a one-dimensional plasma*, Phys. Rev. Lett., 34 (1975), pp. 1083–1085.

[15] J. F. van Diejen and L. Vinet, eds., *Calogero-Moser-Sutherland Models*, CRM Ser. Math. Phys., Springer-Verlag, New York, 2000.

[16] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, UK, 1927.

SIAM J. MATH. ANAL.
Vol. 34, No. 4, pp. 759–773

ON THE EIGENVALUES OF ZAKHAROV–SHABAT SYSTEMS*

M. KLAUS† AND J. K. SHAW†

Abstract. We consider the eigenvalues of a Zakharov–Shabat system on the real line with a complex-valued L^1 potential and show that $\pi/2$ is the threshold L^1 norm of the potential for the formation of eigenvalues. We obtain best possible lower bounds on the number of eigenvalues for a real potential of one sign. This lower bound is exact for the class of single lobe potentials, that is, positive potentials with a single local maximum that is also global.

Key words. Zakharov–Shabat systems, eigenvalue thresholds, number of eigenvalues

AMS subject classifications. Primary, 34L15; Secondary 34L25

PII. S0036141002403067

1. Introduction. Zakharov–Shabat (ZS) systems are non-self-adjoint coupled differential equations of the form [1, 2, 3, 4]

$$v_1' = -i\xi v_1 + q(t)v_2, \quad v_2' = i\xi v_2 - q(t)^* v_1, \tag{1.1}$$

where ξ is a complex-valued eigenvalue (EV) parameter, $q(t)$ is a locally integrable function of the real-variable t, $-\infty < t < \infty$, and the asterisk denotes the complex conjugate. ZS systems have their origin in inverse scattering theory for the nonlinear Schrödinger equation, which in normalized form is [1]

$$iu_z + (1/2)u_{tt} + |u|^2 u = 0, \quad u = u(z,t). \tag{1.2}$$

In (1.2) we can think of $u(z,t)$ as the slowly varying field of a light pulse propagating in an optical fiber under the influence of chromatic dispersion and material nonlinearity [5]; t is normalized local pulse time and z is normalized length along the fiber. The connection between (1.1) and (1.2) is that (1.2) can be formally solved by setting $q(t) = u(0,t)$ and finding $u(z,t)$ by the inverse scattering procedure associated with (1.1) [6, 1, 2]. A further connection is that the EVs of (1.1) correspond to optical solitons of (1.2), that is, particular solutions of (1.2) whose amplitudes are either constant or periodic in z [6, 2]. Both fundamental (constant amplitude) and higher order (periodic) solitons have a functional form involving hyperbolic secants. The EVs are defined as those complex numbers ξ, $\mathrm{Im}(\xi) > 0$, for which (1.1) has a solution $\vec{v}(t) = \binom{v_1(t)}{v_2(t)}$ of integrable square on the real line; that is,

$$\int_{-\infty}^{\infty} (|v_1(t)|^2 + |v_2(t)|^2)\, dt < \infty. \tag{1.3}$$

The initial shape $u(0,t) = q(t)$ does not have to be a hyperbolic secant in order for (1.2) to support solitons, at least in the asymptotic sense. If the initial pulse has enough energy, as measured by the integral

$$E = \int_{-\infty}^{\infty} |q(t)|\, dt, \tag{1.4}$$

*Received by the editors February 22, 2002; accepted for publication (in revised form) July 30, 2002; published electronically February 6, 2003.
http://www.siam.org/journals/sima/34-4/40306.html
†Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0123 (klaus@math.vt.edu, shaw@math.vt.edu).

then solitons will evolve, as components of $u(z,t)$, with increasing z [2]. Specifically, if E is sufficiently large for (1.1) to have EVs, then those EVs completely determine the asymptotic behavior of $u(z,t)$ as $z \to \infty$ [1, 2].

For conventional real, symmetric, and monomodal pulse shapes $q(t)$ in (1.1), such as Gaussians, hyperbolic secants, and rectangles, the EVs are purely imaginary. This is significant because EVs with the same real part can combine to form a higher order, periodic soliton [2]. The property of having purely imaginary EVs has actually been claimed for many years as a general property of real symmetric potentials but is false [3, 4]. The present authors [4] have established that "single lobe" potentials $q(t)$, functions which increase until t reaches a central concentration point and then decrease, do have purely imaginary EVs only. Single lobe potentials will also be important in this paper.

Specifically, in computable cases ([7, 8]; see [3, 4] for additional references) EVs appear when E in (1.4) exceeds certain threshold levels. For example, if $q(t)$ is a positive constant on a compact interval and 0 elsewhere (rectangular pulse), then successive EVs appear as E crosses the levels $E = (2n-1)\pi/2$ $(n = 1, 2, 3, \dots)$ [7]. Similarly, $\pi/2$ is the threshold for potentials which are constant multiples of $\operatorname{sech}(t)$ [8]. In general, the best known bound on $q(t)$ which rules out EVs is $E < 1.32$ [2] ($E < 0.904$ in [1]). There seem to be no general lower bounds on E in the literature which guarantee the existence of EVs.

Let us state our principal results. Unless stated specifically otherwise, we will assume throughout that

$$q \text{ is real-valued and } q \in L^1(-\infty, \infty); \tag{1.5}$$

that is, $E < \infty$ in (1.4). We are going to improve the 1.32 bound to $\pi/2$ and show that $\pi/2$ is best possible in a wider sense than is provided by the computable examples cited above. In section 3 we prove that if $E \le \pi/2$, then there are no purely imaginary EVs of (1.1). Later, in section 4 we strengthen this to claim that there are no EVs at all for complex potentials satisfying $E \le \pi/2$. Define

$$I = \int_{-\infty}^{\infty} q(t)\, dt. \tag{1.6}$$

In section 3 we show for real potentials $q(t)$ that there are at least N purely imaginary EVs, where N is the largest nonnegative integer such that $(2N-1)\pi/2 < |I|$. For $N = 1$ this reduces, of course, to $|I| > \pi/2$, rendering $\pi/2$ as the EV formation threshold for essentially all physically interesting potentials. For single lobe potentials we will prove that there are exactly N EVs if N is defined as above. For single sign potentials (section 4) we show that the largest magnitude of an EV on the imaginary axis strictly dominates the imaginary part of any other EV in the complex plane. In particular, if there are no imaginary EVs, then there are no EVs at all. Moreover, if $|q|$ (for $q(t)$ complex-valued) has no EVs on the imaginary axis, then q has no EVs at all.

In section 2 we establish most of our results on imaginary EVs in the case where q has compact support, because this situation is concrete and sufficiently interesting, physically speaking, to stand on its own. (All real optical pulses have compact support.) Since potentials of noncompact support, such as hyperbolic secants, are also important, we consider noncompact support in section 3; this will be independent of section 2. We treat single sign and complex-valued potentials in section 4. In section 5

we summarize the paper and discuss some applications giving practical criteria under which initial pulses induce solitons.

We want to clarify that E in (1.4) is not technically the energy contained in the initial pulse $q(t) = u(0,t)$; the energy is given instead by the integral of $|q|^2$. However, (1.4) can be still regarded as an indicator. We note, moreover, that theoretical solitons require conservation of energy and therefore do not actually exist in real fibers. Soliton based fiber communication systems use the soliton effect, whereby a launched pulse that is nearly a perfect soliton gradually degrades due to fiber attenuation which can be as low as 5% per kilometer [5].

2. Compact support case. In this section we suppose in addition to (1.5) that $q(t) = 0$ outside an interval $[-d, d]$, $d > 0$. In the scattering theory of (1.1) one defines the Jost solutions $\vec{\psi}(t,\xi)$ and $\vec{\varphi}(t,\xi)$ by the asymptotic properties [1]

$$\vec{\psi}(t,\xi) \cong \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^{i\xi t}, \quad t \to \infty, \quad \vec{\varphi}(t,\xi) \cong \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^{-i\xi t}, \quad t \to -\infty,$$

which guarantee exponentially small solutions, unique up to constant multiples, at $\pm\infty$ for $\operatorname{Im}(\xi) > 0$. Eigenfunctions of (1.1) are multiples of both Jost solutions. Outside $[-d, d]$, where $q(t) = 0$, (1.1) can be solved in closed form and the EV condition (1.3) gives

$$v_1(-d) = 1, \quad v_2(-d) = 0, \quad v_1(d) = 0, \tag{2.1}$$

where the first of these is actually a normalization and the last is a condition for the existence of an EV. That is, the first two of (2.1) hold for all ξ and the last holds when ξ is an EV.

Since the objects of our investigation are the purely imaginary EVs of (1.1), we set

$$\xi = is, \quad s \geq 0.$$

We employ a Prüfer transformation in (1.1),

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} \rho\cos\theta \\ \rho\sin\theta \end{pmatrix}, \tag{2.2}$$

so that $\rho = \rho(t;s)$ and $\theta = \theta(t;s)$ satisfy

$$\theta' = -q(t) - s\,\sin(2\theta) \tag{2.3}$$

and

$$\rho' = s\,\rho\cos(2\theta), \tag{2.4}$$

along with the initial conditions $\theta(-d;s) = 0$, $\rho(-d;s) = 1$ (by (2.1)). The primes in (2.3) and (2.4) denote differentiation with respect to t. The relevant equation for us will be (2.3) because $\theta(t;s)$ contains all the important information about the location of EVs on the imaginary axis. Note that $\xi = is$ is an EV if and only if

$$\theta(d;s) = \frac{(2k-1)\pi}{2}$$

for some integer k.

LEMMA 2.1. *We have* $\theta(d;s) \to 0$ *as* $s \to \infty$.

Proof. We write (2.3) in the form

$$\theta' + 2s\theta = -q(t) - s[\sin(2\theta) - 2\theta]$$

and convert it to an integral equation

$$\theta(t;s) = -e^{-2st}\int_{-d}^{t} e^{2s\tau} q(\tau)\,d\tau - se^{-2st}\int_{-d}^{t} e^{2s\tau}[\sin(2\theta(\tau;s)) - 2\theta(\tau;s)]\,d\tau, \tag{2.5}$$

where the initial condition $\theta(-d;s) = 0$ has been taken into account. Let

$$\eta(s) = \sup_{-d\le t\le d}\left(e^{-2st}\int_{-d}^{t} e^{2s\tau}|q(\tau)|\,d\tau\right)$$

and note that $\eta(s) \to 0$ as $s \to \infty$. Let $\varepsilon > 0$ and suppose without loss of generality that $\varepsilon < 1/2$. Choose s_0 so large that

$$\eta(s) < 2\varepsilon/3, \quad s > s_0. \tag{2.6}$$

We will show that this implies $\sup_{-d\le t\le d}|\theta(t;s)| < \varepsilon$. Fix $s > s_0$ and let

$$m(t;s) = \sup_{-d\le\tau\le t}|\theta(\tau;s)|.$$

Using the inequality $|\sin(z) - z| \le |z|^3/6$, for real z, the estimate

$$s\int_{-d}^{t} e^{-2s(t-\tau)}\,d\tau \le 1/2$$

in (2.5) implies

$$m(t;s) < \frac{2\varepsilon}{3} + \frac{2}{3}m(t;s)^3. \tag{2.7}$$

The function $f(x) = x - (2/3)x^3 - (2\varepsilon/3)$ has two positive roots r_1 and r_2, $0 < r_1 < (1/\sqrt{2}) < r_2$; note that $f(x)$ has a positive maximum at $x = 1/\sqrt{2}$ because $\varepsilon < (1/2) < 1/\sqrt{2}$. Inequality (2.7) says that $f(m(t;s)) < 0$. Since $m(-d;s) = 0$ and $m(t;s)$ is continuous in t, it follows that $m(t;s) \le r_1$, and so $m(t;s) \le 1/\sqrt{2}$. Substituting $m(t;s) \le 1/\sqrt{2}$ into the right side of (2.7) yields

$$m(t;s) < \frac{2\varepsilon}{3} + \frac{1}{3}m(t;s), \quad -d \le t \le d,$$

which is to say $m(t;s) < \varepsilon$. Then $|\theta(t;s)| < \varepsilon$ and the proof is complete.

THEOREM 2.2. *Let* N *be the largest nonnegative integer such that* $(2N-1)\pi/2 < |I|$, *where* I *is given by* (1.6) *(integrated over* $[-d,d]$*). Then* (1.1) *has at least* N *purely imaginary EVs. In particular, if* $|I| > \pi/2$, *then there is at least one purely imaginary EV.*

Proof. For $s = 0$ (2.3) implies that $I = -\theta(d;0)$, and so $|\theta(d;0)| > (2N-1)\pi/2$. By continuity of $\theta(d;s)$ and Lemma 2.1, there must be N values $0 < s_1 < s_2 < \cdots < s_N$ such that $|\theta(d;s_k)| = (2(N-k)+1)\pi/2$, meaning that each $\xi_k = is_k$ is an EV.

Define

$$q_+(t) = \max[q(t),0], q_-(t) = \max[-q(t),0]$$

so that $q = q_+ - q_-$.

We will need the following result.

COMPARISON THEOREM (see [9, p. 122]). *Let the function $f(t, y)$ satisfy a local Lipschitz condition in y and define the operator P by $P(g) = g' - f(t, g)$. Let g_1 and g_2 be absolutely continuous functions on $[t_1, t_2]$ such that $g_1(t_1) \le g_2(t_1)$ and $P(g_1) \le P(g_2)$ almost everywhere on $[t_1, t_2]$. Then either $g_1(t) < g_2(t)$ everywhere in $[t_1, t_2]$ or there exists a point c, $t_1 \le c \le t_2$, such that $g_1(t) = g_2(t)$ in $[t_1, c]$ and $g_1(t) < g_2(t)$ in $(c, t_2]$.*

THEOREM 2.3. *Suppose that*

$$\int_{-d}^{d} q_+(\tau)\, d\tau \le \pi/2 \quad \text{and} \quad \int_{-d}^{d} q_-(\tau)\, d\tau \le \pi/2. \tag{2.8}$$

Then there are no imaginary EVs. In particular, this is true if $\int_{-d}^{d} |q(\tau)|\, d\tau \le \pi/2$.

Proof. We will show that $|\theta(d; s)| < \pi/2$ for $s > 0$. Note that when $s = 0$, (2.3) and (2.8) imply that $|\theta(t, 0)| \le \pi/2$, $-d \le t \le d$. Let $w_\pm(t; s)$ be the solutions of

$$w'_\pm = \pm q_\mp(t) - s \sin[2w_\pm], \quad w_\pm(-d; s) = 0. \tag{2.9}$$

For application of the comparison theorem cited above let $f(t, g) = -q(t) - s \sin[2g(t)]$. Then $P(g) = g' + q(t) + s \sin[2g(t)]$ and so $P(\theta) = 0$. Furthermore,

$$\begin{aligned} P(w_-) &= w'_- + q(t) + s \sin[2w_-] \\ &= w'_- + q_+(t) - q_-(t) + s \sin[2w_-] \\ &= w'_- + q_+(t) + s \sin[2w_-] - q_-(t) \\ &= 0 - q_-(t) \le 0, \end{aligned}$$

and we conclude that $w_-(t; s) \le \theta(t; s)$ for $-d \le t \le d$, $s > 0$. Analogously, one shows $\theta(t; s) \le w_+(t; s)$. Since the initial value problem $z' = -s \sin(2z)$, $z(-d) = 0$, has the unique solution $z = 0$, then we have $w_-(t; s) \le 0 \le w_+(t; s)$, also by the comparison theorem. By (2.9) and the second condition in (2.8),

$$w_+(t; s) \le \int_{-d}^{t} q_-(\tau)\, d\tau, \quad -d \le t \le d. \tag{2.10}$$

Unless $q_-(t) \equiv 0$, then strict inequality holds in (2.10); that is,

$$w_+(d; s) < \int_{-d}^{d} q_-(\tau)\, d\tau. \tag{2.11}$$

For otherwise if equality held, then it would follow that

$$w_+(t; s) = \int_{-d}^{t} q_-(\tau)\, d\tau, \quad -d \le t \le d,$$

by the comparison theorem, and thus $w'_+(t; s) = q_-(t)$. Therefore $\sin[2w_+(t; s)] \equiv 0$, $w_+(t; s) \equiv 0$, and $q_-(t) \equiv 0$, a contradiction. So (2.11) and (2.8) together imply $\theta(d; s) \le w_+(t; s) < \pi/2$ for $s > 0$. If $q_-(t) \equiv 0$, then $w_+(t; s) \equiv 0$ and $\theta(d; s) < \pi/2$ is obviously true. Similarly, using $w_-(t; s)$ we deduce that $\theta(d; s) > -\pi/2$, $s > 0$, and so $|\theta(d; s)| < \pi/2$, $s > 0$, in all cases. Therefore purely imaginary EVs do not exist.

We note that the true number of EVs may be strictly larger than N due to the fact that $\theta(d;s)$ need not be monotone decreasing in general. That is, there can be multiple crossing events $\theta(d;s) = (2k-1)\pi/2$ without monotonicity. As an example, let $q(t) = h$ for $1 \le |t| \le 2$ and $q(t) = 0$ otherwise. As h increases, the first EV appears at $\xi = 0$ for $h = \pi/4$ (so that $I = \pi/2$ in (1.6)) and moves up the imaginary axis. According to Theorem 2.2 there is at least one imaginary EV when $\pi/4 < h < 3\pi/4$. However, a detailed analysis of this example shows that for $h = 2.2$ there are three purely imaginary EVs at approximately $\xi = 0.28i$, $0.63i$, and $1.03i$. The reason is that as $h \to h_c = 2.178$ (approximately) a pair of EVs having opposite nonzero real parts converge to and collide on the imaginary axis at about $\xi = 0.44i$. After the collision the two complex EVs become a pair of purely imaginary EVs so that there are a total of three imaginary EVs for h slightly larger than h_c. The occurrence of EVs colliding on the imaginary axis means that $\theta(d;s)$ need not be monotone in general. In this example we have with $h = 2.2$ that $\theta(2;s_1) = \theta(2;s_2) = -3\pi/2$ at $s_1 = 0.28$ and $s_2 = 0.63$, approximately.

We now show that for single lobe potentials, as mentioned in the introduction, Theorem 2.2 can be strengthened to assert that there are exactly N EVs. To be specific, $q(t)$ is called a single lobe potential if q satisfies (1.5), is bounded, piecewise smooth, and nondecreasing to the left of $t = 0$ and nonincreasing to the right of $t = 0$. By piecewise smooth we mean that $q(t)$ and $q'(t)$ have left- and right-hand limits for all t and that in any bounded interval $q(t)$ has at most finitely many jump discontinuities. Since (1.1) is invariant under shifts in the t variable, there is no loss of generality in taking $t = 0$ to be the point of energy concentration. Our results for single lobe potentials are also true if $q(t) < 0$ and $-q(t)$ is single lobe.

THEOREM 2.4. *In addition to the hypotheses of Theorem* 2.2 *assume that* $q(t)$ *is single lobe. Then there are exactly* N *purely imaginary EVs and no nonimaginary EVs.*

Proof. The statement claiming no nonimaginary EVs is proved in [4]. Putting $\xi = is$ in (1.1) gives the system

$$v_1' = sv_1 + q(t)v_2, \quad v_2' = -sv_2 - q(t)v_1. \tag{2.12}$$

We will show that multiple crossing events $\theta(d;s) = (2k-1)\pi/2$ are ruled out by the condition $\frac{d}{ds}\theta(d;s) > 0$, when s is an EV, for single lobe potentials. To this end, differentiate (2.12) with respect to s, using an overdot to denote the s derivative, to obtain

$$\dot{v}_1' = v_1 + s\dot{v}_1 + q(t)\dot{v}_2, \quad \dot{v}_2' = -v_2 - s\dot{v}_2 - q(t)\dot{v}_1. \tag{2.13}$$

Substitute (2.13) into $(\dot{v}_1 v_2 - v_1\dot{v}_2)'$, where the prime denotes differentiation with respect to t, expand, and simplify to obtain $(\dot{v}_1 v_2 - v_1\dot{v}_2)' = 2v_1 v_2$. The conditions (2.1) yield $\dot{v}_1(-d) = \dot{v}_2(-d) = 0$, and following integration we obtain

$$(\dot{v}_1 v_2 - v_1\dot{v}_2)(d) = 2\int_{-d}^{d} v_1(\tau)v_2(\tau)\,d\tau. \tag{2.14}$$

Noting (2.2) the quotient rule derivative (with respect to s) of $(v_2/v_1) = \tan(\theta)$ gives by (2.14)

$$\dot{\theta}(d;s) = \frac{-2}{v_1^2(d) + v_2^2(d)} \int_{-d}^{d} v_1(\tau)v_2(\tau)\,d\tau, \tag{2.15}$$

where $v_1(d) = 0$ if s is an EV. In [4] the authors have shown that the integral in (2.15) is negative for single lobe potentials. For completeness we give the gist of the argument, which comes from multiplying the first of (2.12) by v_1, solving for $v_1 v_2$, and integrating over $[0, d]$ to obtain

$$\int_0^d v_1(\tau)v_2(\tau)\,d\tau = -\frac{v_1^2(0)}{2q(0)} + \frac{1}{2}\int_0^d \frac{v_1^2(\tau)q'(\tau)}{q^2(\tau)}\,d\tau - s\int_0^d \frac{v_1^2(\tau)}{q(\tau)}\,d\tau. \tag{2.16}$$

Convergence of the integrals on the right of (2.16) can be justified for a single lobe $q(t)$. Since $q'(t) \le 0$ for $t \ge 0$, the right side of (2.16) is negative. Working with the other term in (2.12) shows similarly that the corresponding integral over $[-d, 0]$ is also negative. Therefore $\dot\theta(d; s) > 0$ at EV crossings, and the proof of Theorem 2.2 shows that there are exactly N purely imaginary EVs.

3. Noncompact support case. The Prüfer transformation (2.2) can also be applied if the potential $q(t)$ does not have compact support. The extension of the Prüfer method to the full line may be of independent interest. In this section we assume (1.5) only.

Equation (2.3) is still the most significant, but now the solution is required to satisfy

$$\operatorname*{Lim}_{t\to-\infty}\ \theta(t; s) = 0. \tag{3.1}$$

The integral equation corresponding to (2.5) now reads

$$\theta(t; s) = -e^{-2st}\int_{-\infty}^t e^{2s\tau}q(\tau)\,d\tau - se^{-2st}\int_{-\infty}^t e^{2s\tau}[\sin(2\theta(\tau; s)) - 2\theta(\tau; s)]\,d\tau. \tag{3.2}$$

Equation (3.2) can be solved by iteration, and it follows by standard arguments that (3.2) is the unique solution to (2.3) on $-\infty < t < \infty$ satisfying (3.1). Our goal now is to describe the behavior of $\theta(t; s)$ as $t \to \infty$.

LEMMA 3.1. *For each integer k and for each $s \ge 0$ there is a unique solution $\varphi_k(t; s)$ of* (2.3) *such that*

$$\operatorname*{Lim}_{t\to\infty}\ \varphi_k(t; s) = \frac{(2k-1)\pi}{2} \tag{3.3}$$

uniformly in s.

Proof. It suffices to prove the lemma for $k = 0$, in view of the fact that $\varphi_k(t; s) = \varphi_0(t; s) + k\pi$, which may be verified by substitution in (2.3). To this end define $\chi_0(t; s)$ as the unique solution satisfying

$$\chi_0'(t; s) = q(-t) - s\,\sin[2\chi_0(t; s)], \quad \chi_0(t; s) \to 0 \text{ as } t \to -\infty.$$

This is the same differential equation as (2.3) but with $q(t)$ replaced by $-q(-t)$. Then

$$\varphi_0(t; s) = \chi_0(-t; s) - \pi/2$$

satisfies the first conclusion of the lemma with $k = 0$. To prove the uniformity in s we use the integral equation for $\chi_0(t; s)$ and mimic the proof of Lemma 2.1. Given $\varepsilon > 0$ we can find a number T_ε such that

$$\left|e^{-2st}\int_{-\infty}^t e^{2s\tau}q(-\tau)\,d\tau\right| \le \int_{-\infty}^{-T_\varepsilon} |q(-\tau)|\,d\tau \le 2\varepsilon/3$$

for $t < -T_\varepsilon$ and $s \geq 0$. Proceeding as in the proof of Lemma 2.1 we conclude that $|\chi_0(t;s)| \leq \varepsilon$ for $t < -T_\varepsilon$ and $s \geq 0$. Hence (3.3) is valid uniformly in s.

Note that $\xi = is$ is an EV if and only if $\theta(t;s)$ of (3.1) and (3.2) is a multiple of one of the functions $\varphi_k(t;s)$.

THEOREM 3.2. *Suppose that $q(t)$ satisfies* (1.5) *and that $y(t;s)$ is any solution of* (2.3) *with $s > 0$. Then the limit*

$$L_y(s) = \operatorname*{Lim}_{t \to \infty} y(t;s) \tag{3.4}$$

exists and, moreover, $L_y(s) = k\pi/2$ for some integer k.

Proof. The idea is to exploit the fact that the term $s\sin[2y(t;s)]$ dominates the right side of (2.3) when t is large. Note that if $q(t) = 0$, then (2.3) has the constant solutions $y(t;s) = k\pi/2$, with k an integer.

If $y(0;s) = \varphi_k(0;s)$ for some k, then $y(t;s) = \varphi_k(t;s)$ by uniqueness of solutions and the conclusion follows from Lemma 3.1. Thus we can suppose that $y(0;s) \neq \varphi_k(0;s)$ for any k. Suppose first that $\varphi_0(0;s) < y(0;s) < \varphi_1(0;s)$. We will show that $y(t;s) \to 0$, $t \to \infty$. By the uniqueness part of Lemma 3.1, $y(t;s)$ cannot converge to either $\pm\pi/2$ (by uniqueness of the φ_k). Suppose that $y(t;s)$ does not converge to 0. Then there is a $\delta > 0$ and a sequence $t_n \to \infty$ such that for all n either

$$\delta < y(t_n;s) < (\pi/2) - \delta \quad \text{or} \quad -(\pi/2) + \delta < y(t_n;s) < -\delta.$$

Without loss of generality, assume the former. We will derive a contradiction. Pick N so large that

$$\int_{t_N}^{\infty} |q(\tau)|\, d\tau < \delta/2 \tag{3.5}$$

and note that on the interval $(\delta/2) < z < (\pi - \delta)/2$ we have

$$\sin(2z) > c_\delta \tag{3.6}$$

for some $c_\delta > 0$. Using (3.6) in (2.3) we obtain

$$y'(t;s) \leq -q(t) - c_\delta s$$

so long as $(\delta/2) < y(t;s) < (\pi - \delta)/2$. Therefore, if $y(t;s)$ satisfies these bounds for $\alpha \leq t \leq \beta$, then

$$y(t;s) \leq y(\alpha;s) - c_\delta s(t-\alpha) - \int_\alpha^t q(\tau)\, d\tau, \quad \alpha \leq t \leq \beta. \tag{3.7}$$

We first claim that $(\delta/2) < y(t;s) < (\pi - \delta)/2$ for all $t \geq t_N$. Otherwise, one of the following situations must occur:

(a) there is an interval $[\alpha, \beta]$ with $\alpha > t_N$ such that $y(\alpha;s) = (\pi/2) - \delta < y(t;s) < (\pi - \delta)/2$, for $\alpha < t < \beta$, and $y(\beta;s) = (\pi - \delta)/2$;

(b) there is an interval $[\alpha, \beta]$ with $\alpha > t_N$ such that $y(\alpha;s) = \delta/2 < y(t;s) < \delta$, for $\alpha < t < \beta$, and $y(\beta;s) = \delta$.

In case (a), (3.7) implies $y(t;s) < (\pi/2) - \delta + (\delta/2) = (\pi - \delta)/2$, for $\alpha \leq t \leq \beta$, and this contradicts $y(\beta;s) = (\pi - \delta)/2$. In case (b), we conclude that $y(t;s) < (\delta/2) + (\delta/2) = \delta$, contradicting $y(\beta;s) = \delta$. Thus $(\delta/2) < y(t;s) < (\pi - \delta)/2$ for all $t \geq t_N$. However, then (3.7) leads to an obvious contradiction since it says that $y(t;s)$ becomes negative

for large enough t since $s > 0$. Similarly, one argues that there are no sequences t_n such that $-(\pi/2) + \delta < y(t_n; s) < -\delta$. Therefore $y(t; s) \to 0$.

For solutions $y(t; s)$ satisfying $\varphi_k(0; s) < y(0; s) < \varphi_{k+1}(0; s)$ we use the fact that $\tilde{y}(t; s) = y(t; s) - k\pi$ satisfies $\varphi_0(0; s) < \tilde{y}(0; s) < \varphi_1(0; s)$ so that $\tilde{y}(t; s) \to 0$ as $t \to \infty$; i.e., $y(t; s) \to k\pi$ as $t \to \infty$. This completes the proof of Theorem 3.2. Note that the φ_k have limits $L_{\varphi_k}(s) = (2k-1)\pi/2$ which are odd multiples of $\pi/2$, while all other solutions $y(t; s)$ have limits which are even multiples of $\pi/2$.

THEOREM 3.3. *The conclusions of Theorems* 2.2 *and* 2.3 *hold for general* $q(t)$ *satisfying* (1.5).

Proof. We prove only the extension of Theorem 2.2; the proof of the extension of Theorem 2.3 is similar. For the extension of Theorem 2.2 it suffices to show that $L_\theta(s) \to 0$, $s \to \infty$, without skipping any odd multiples of $\pi/2$. That is, if $|L_\theta(0)| > \frac{(2N-1)\pi}{2}$, we will show that for $k = 1, 2, \ldots, N$ there are s_k such that $|L_\theta(s_k)| = \frac{(2(N-k)+1)\pi}{2}$. Note that $\theta(t; 0) = -\int_{-\infty}^{t} q(\tau)\, d\tau$ and so $L_\theta(0) = -I$ (from (1.6)).

In showing $L_\theta(s) \to 0$ we will actually prove that $L_\theta(s) = 0$ for $s > \tilde{s}$, with $\tilde{s}$ sufficiently large. Let

$$\delta = |L_\theta(0)| - \frac{(2N-1)\pi}{2},$$

where N is defined as in Theorem 2.2. Note that $0 < \delta \leq \pi$ by the definition of N. Pick $\tilde{t}$ so large that $|\varphi_0(\tilde{t}; s) + (\pi/2)| < (\delta/4)$, $|\varphi_1(\tilde{t}; s) - (\pi/2)| < (\delta/4)$, for $s \geq 0$, and $|\theta(\tilde{t}; 0)| > \frac{(2N-1)\pi}{2} + (\delta/4)$. By a slight extension of Lemma 2.1 to the case $d = \infty$ we can conclude that there is an $\tilde{s} > 0$ such that $|\theta(\tilde{t}; s)| < \pi/4$ for $s > \tilde{s}$. Therefore $\varphi_0(\tilde{t}; s) < \theta(\tilde{t}; s) < \varphi_1(\tilde{t}; s)$ if $s > \tilde{s}$. As in the proof of Theorem 3.2 it follows that $\theta(t; s) \to 0$, $t \to \infty$; that is, $L_\theta(s) = 0$, $s > \tilde{s}$.

Suppose now that $L_\theta(0) > \frac{(2N-1)\pi}{2}$; the case $L_\theta(0) < -\frac{(2N-1)\pi}{2}$ is handled similarly. Since $\theta(t; s)$ is continuous in s there are at least N crossings s_k, where $\theta(\tilde{t}; s_k) = \varphi_{N-k+1}(\tilde{t}; s_k)$, so that $\theta(t; s_k) = \varphi_{N-k+1}(t; s_k)$ for all t. Each s_k corresponds to an imaginary EV, which proves the extension of Theorem 2.2.

THEOREM 3.4. *Suppose* $q(t)$ *satisfies* (1.5) *and is an odd function. Then there are no imaginary EVs.*

Proof. Suppose $\xi = is$ is an EV with eigenfunction $\theta(t; s)$ such that $L_\theta(s) = \frac{(2m-1)\pi}{2}$ for some integer m. By substitution $w(t; s) = \theta(-t; s) + \frac{(2m-1)\pi}{2}$ is also a solution of (2.3) since $q(t)$ is odd. Since $\theta(-t; s) \to 0$, $t \to \infty$, then $\mathrm{Lim}_{t\to\infty} w(t; s) = \frac{(2m-1)\pi}{2} = L_\theta(s)$ and therefore $w(t; s) \equiv \theta(t; s)$. Since $w(0; s) = \theta(0; s) + \frac{(2m-1)}{2} \neq \theta(0; s)$, we have a contradiction.

We close the section with an alternate proof of Theorem 3.4. Let $\xi = is$ be an EV with eigenfunction $\vec{v}(t)$. By the odd symmetry of $q(t)$ the function $\eta(t) = \binom{-v_2(-t)}{v_1(-t)}$ is an eigenfunction. Thus $\vec{\eta} = K\vec{v}$ for some constant K. Then $-v_2(0) = Kv_1(0)$ and $v_1(0) = Kv_2(0)$, neither of which is possible unless $\vec{v}(t) \equiv 0$.

4. More general potentials. In this section we suppose that

$$q \in L^1(-\infty, \infty) \tag{4.1}$$

and allow $q(t)$ to be nonreal for some of our results. We begin with the observation that if ξ ($\mathrm{Im}(\xi) > 0$) is an EV for (1.1), then the eigenfunction $\vec{v}(t)$ satisfies

$$\begin{aligned} v_1(t) &= -e^{-i\xi t}\int_t^\infty e^{i\xi\tau}q(\tau)v_2(\tau)\,d\tau, \\ v_2(t) &= -e^{i\xi t}\int_{-\infty}^t e^{-i\xi\tau}q(\tau)^*v_1(\tau)\,d\tau. \end{aligned} \tag{4.2}$$

Equation (4.2) is obtained by converting (1.1) to a system of integral equations using the fact that $v_1, v_2 \in L^2(-\infty,\infty)$. It is convenient to recast (4.2) as an EV problem for a compact operator which will play a role analogous to that of the Birman–Schwinger kernel [10, p. 98] for the Schrödinger equation. To this end we put

$$q(t) = |q(t)|e^{i\sigma(t)}, \quad -\pi < \sigma(t) \le \pi,$$

and let

$$U(t) = \begin{pmatrix} e^{-i\sigma(t)} & 0 \\ 0 & e^{i\sigma(t)} \end{pmatrix}. \tag{4.3}$$

We also introduce the matrix integral kernel

$$A_\xi(t,\tau) = \begin{pmatrix} e^{i\xi(\tau-t)}H(\tau-t) & 0 \\ 0 & e^{i\xi(t-\tau)}H(t-\tau) \end{pmatrix}, \tag{4.4}$$

where H denotes the Heaviside step function, and we put

$$\vec{w} = |q|^{1/2}\vec{v}, \quad J = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{4.5}$$

Then (4.2) can be written as

$$\vec{w}(t) = \int_{-\infty}^\infty K_\xi(t,\tau)\vec{w}(\tau)\,d\tau, \tag{4.6}$$

where

$$K_\xi(t,\tau) = -|q(t)|^{1/2}A_\xi(t,\tau)J|q(\tau)|^{1/2}U(\tau). \tag{4.7}$$

We denote by K_ξ the integral operator induced by the kernel $K_\xi(t,\tau)$. Note that every EV ξ of (1.1) with corresponding eigenfunction $\vec{v}$ gives rise to an eigenfunction $\vec{w}$ for the EV 1 of K_ξ. Conversely, if $\vec{w}(t)$ is a square integrable solution of (4.6), then

$$\vec{v}(t) = \int_{-\infty}^\infty A_\xi(t,\tau)J|q(\tau)|^{1/2}U(\tau)\vec{w}(\tau)\,d\tau$$

is an eigenfunction of (1.1) for the EV ξ. This connection is known as the Birman–Schwinger principle.

LEMMA 4.1. *Suppose q satisfies* (4.1) *and let* $\mathrm{Im}(\xi) \ge 0$. *Then K_ξ is Hilbert–Schmidt. If q is real and of one sign, then K_ξ is unitarily equivalent to $-K_\xi$, and if $\xi = is$ $(s > 0)$, then K_{is} is self-adjoint. Moreover, $\|K_{is}\| \to 0$, $s \to \infty$.*

Proof. The Hilbert–Schmidt norm of K_ξ is given by

$$\|K_\xi\|^2_{H.S.} = \iint_{\Re^2} Tr[K^+_\xi(t,\tau)K_\xi(t,\tau)]\,dt\,d\tau,$$

where the superscript "+" denotes the matrix adjoint and "Tr" the matrix trace. From (4.4) and (4.7) we obtain (putting $\mathrm{Im}(\xi) = s$)

$$\|K_\xi\|^2_{H.S.} = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{t}|q(\tau)|e^{2s\tau}\,d\tau\right)|q(t)|e^{-2st}\,dt \le \left(\int_{-\infty}^{\infty}|q(t)|\,dt\right)^2 < \infty \tag{4.8}$$

by (4.1), proving the first part of the lemma. If q is real and of one sign, then $\sigma(t) = 0$ ($\sigma(t) = \pi$) if $q(t) \ge 0$ ($q(t) \le 0$), and hence by (4.3) $U(t) = \pm I$, where I is the identity matrix. Let $S = \left(\begin{smallmatrix}1 & 0\\ 0 & -1\end{smallmatrix}\right)$. Then $SK_\xi S^{-1} = -K_\xi$, proving that K_ξ is unitarily equivalent to $-K_\xi$. Using $U(t) = \pm I$, $\xi = is$ in (4.7) and the fact that $A^*_{is} = JA_{is}J$, we conclude that $K^*_{is} = K_{is}$. The last statement follows from (4.8) and the Lebesgue dominated convergence theorem.

From (4.8) it follows that $\|K_\xi\|_{H.S.} < \|q\|_{L^1}$ if $\mathrm{Im}(\xi) = s > 0$. Hence there can be no EVs if $\|q\|_{L^1} \le 1$. However, $\|q\|_{L^1} \le 1$ is not as strong as the bound $\|q\|_{L^1} \le 1.32$ mentioned in the introduction, and the latter is not best possible. The optimal L^1 bound on q that guarantees the absence of EVs is obtained in the next theorem.

THEOREM 4.2. *Suppose that q is complex-valued and satisfies* (4.1). *If*

$$\int_{-\infty}^{\infty}|q| \le \pi/2, \tag{4.9}$$

then (1.1) *has no EVs.*

Proof. Put $\xi = \mu + is$, $s > 0$, and

$$W_\mu(t) = \begin{pmatrix} e^{-i\mu t} & 0\\ 0 & e^{i\mu t}\end{pmatrix}.$$

Then we can write $K_\xi(t,\tau)$ given in (4.7) as

$$K_\xi(t,\tau) = W_\mu(t)K_{is}(t,\tau)W_\mu(\tau).$$

Here we have used $A_\xi(t,\tau) = W_\mu(t)A_{is}(t,\tau)W^*_\mu(\tau)$ and $W^*_\mu(\tau)J = JW_\mu(\tau)$. The unitarity of W_μ and U implies that

$$\|K_\xi\| = \|K_{is}\| = \|\tilde K_{is}\|, \tag{4.10}$$

where $\tilde K_{is}$ is the self-adjoint operator with kernel

$$\tilde K_{is}(t,\tau) = -|q(t)|^{1/2}A_{is}(t,\tau)J|q(\tau)|^{1/2}.$$

We are going to use the Birman–Schwinger principle for the potential $|q|$. Now (4.9) together with Theorem 3.3 imply that (1.1) with $|q|$ in place of q has no EV on the imaginary axis. Hence $\|\tilde K_{is}\| < 1$, $s > 0$, for if not, then by Lemma 4.1 there would exist a point $s_0 > 0$ such that $\|\tilde K_{is_0}\| = 1$. This would imply that $\tilde K_{is_0}$ has EV 1, which by the Birman–Schwinger principle implies further that $\xi = is_0$ is an EV of (1.1) for the potential $|q|$, a contradiction. Then $\|K_\xi\| < 1$ by (4.10) and K_ξ does not have an EV 1 for any ξ. The conclusion of the theorem follows.

By Theorem 2.2 the bound $\pi/2$ is optimal as claimed.

Note that the ZS system (1.1) for $|q|$ satisfies (4.9) if it has no imaginary EVs, by Theorems 2.3 and 3.3. This proves the following corollary.

COROLLARY 4.3. *Suppose that q is complex-valued and satisfies* (4.1) *and that* (1.1) *with $|q|$ in place of q has no imaginary EVs. Then* (1.1) *for q has no EVs.*

Corollary 4.3 implies that if q is of one sign with a complex EV $\xi_1 = \mu_1 + is_1$ such that $\mu_1 \neq 0$ and $s_1 > 0$, then there must exist a purely imaginary EV $\xi_0 = is_0$. The next theorem says that there is such an EV with $s_0 > s_1$. That is, there is a purely imaginary EV which dominates the magnitudes of the imaginary parts of all other EVs.

THEOREM 4.4. *Suppose that q satisfies* (4.1) *and either $q(t) \geq 0$ or $q(t) \leq 0$. If* (1.1) *has EVs, then there is a purely imaginary EV whose imaginary part is strictly larger than the imaginary part of any other EV.*

Proof. Suppose that $\xi_1 = \mu_1 + is_1$, with $\mu_1 \neq 0$ and $s_1 > 0$, is an EV. Let $\vec{u} = \binom{u_1}{u_2}$ be the normalized eigenfunction for the EV 1 of K_{ξ_1}. Let $\vec{u}_{abs}$ be the vector whose components are $|u_1|$ and $|u_2|$, respectively. Define

$$\begin{aligned} G_{12}(t,\tau) &= -u_1^*|q(t)|^{1/2}e^{i\xi_1(\tau-t)}H(\tau-t)|q(\tau)|^{1/2}e^{-i\sigma(\tau)}u_2(\tau), \\ G_{21}(t,\tau) &= -u_2^*|q(t)|^{1/2}e^{i\xi_1(t-\tau)}H(t-\tau)|q(\tau)|^{1/2}e^{i\sigma(\tau)}u_1(\tau) \end{aligned} \tag{4.11}$$

so that

$$1 = (\vec{u}, K_{\xi_1}\vec{u}) = \iint_{\Re^2} G_{12}(t,\tau)\,dt\,d\tau + \iint_{\Re^2} G_{21}(t,\tau)\,dt\,d\tau$$

and

$$|(\vec{u}_{abs}, \tilde{K}_{is_1}\vec{u}_{abs})| = \iint_{\Re^2} |G_{12}(t,\tau)|\,dt\,d\tau + \iint_{\Re^2} |G_{21}(t,\tau)|\,dt\,d\tau.$$

We obviously have $(\vec{u}, K_{\xi_1}\vec{u}) \leq (\vec{u}_{abs}, \tilde{K}_{is_1}\vec{u}_{abs})$ and equality holds if and only if

$$\iint_{\Re^2} G_{rs}(t,\tau)\,d\tau\,dt = \iint_{\Re^2} |G_{rs}(t,\tau)|\,d\tau\,dt \tag{4.12}$$

for $(r,s) = (1,2)$ or $(2,1)$. Now (4.12) holds if and only if $G_{rs}(t,\tau) = e^{i\beta_{rs}}|G_{rs}(t,\tau)|$ for some real number β_{rs} [11, p. 174]. Putting $u_k(t) = e^{i\varphi_k(t)}|u_k(t)|$ $(k = 1,2)$ we see from (4.11) that (4.12) can hold only if

$$\begin{aligned} e^{-i(\varphi_1(t)+\mu_1 t)}e^{i(\varphi_2(\tau)+\mu_1\tau-\sigma(\tau)+\pi)} &= e^{i\beta_{12}}, \quad t < \tau, \\ e^{-i(\varphi_2(t)-\mu_1 t)}e^{i(\varphi_1(\tau)-\mu_1\tau+\sigma(\tau)+\pi)} &= e^{i\beta_{21}}, \quad t > \tau. \end{aligned}$$

It follows by a separation of variables argument that each of the exponentials on the left side must be constant. However, this is possible only if $\mu_1 = 0$, which contradicts our assumption. We conclude that in (4.12) equality cannot hold; that is, we have

$$1 = (\vec{u}, K_{\xi_1}\vec{u}) < (\vec{u}_{abs}, \tilde{K}_{is_1}\vec{u}_{abs})$$

and therefore $\|\tilde{K}_{is_1}\| > 1$. By Lemma 4.1 there exists $s_0 > s_1$ such that $\|\tilde{K}_{is_0}\| = 1$. In view of (4.10) $\|K_{is_0}\| = 1$ and thus K_{is_0} has EV 1. Then $\xi_0 = is_0$ is an EV of (1.1) whose imaginary part is strictly larger than s_1. It follows that the purely imaginary EV of largest modulus has an imaginary part that is strictly larger than that of any other EV.

5. Summary and discussion. We have established $\pi/2$ as the threshold L^1 norm of the potential (complex-valued) for the formation of EVs in a ZS system. Bounds for the number of imaginary EVs were obtained and shown to be best possible for the class of single lobe potentials. For real, single signed potentials, an imaginary EV of largest magnitude dominates the magnitudes of the imaginary parts of all other EVs, imaginary or not. ZS systems with odd L^1 potentials are free of imaginary EVs.

Theorem 3.3 provides a way to extend criteria for soliton formation found in the current literature. In (1.2) let $q(t) = u(0,t) = N\ \mathrm{sech}(t)$. Since $\int_{-\infty}^{\infty} \mathrm{sech}(t)\,dt = \pi$, then by Theorem 3.3 the condition for a fundamental soliton is $\frac{1}{2} < N < \frac{3}{2}$, which agrees exactly with [5]. This extends to an order K soliton, K a positive integer, when $K - \frac{1}{2} < N < K + \frac{1}{2}$.

The discussion in [5] takes place in the context of normalized, dimensionless units. In physical units the nonlinear Schrödinger equation is

$$iA_z = \frac{\beta_2}{2} A_{TT} - \gamma |A|^2 A, \quad A(0,T) = f(T), \tag{5.1}$$

with $f(T)$ specified, where $A = A(z,T)$ is the "slowly varying" field, z is physical length, T is local pulse physical time, $\beta_2 < 0$ is the fiber dispersion constant, and γ is the material nonlinearity constant [5]. Here $f(T)$ is assumed real-valued. Solitons occur only in the anomalous dispersion case $\beta_2 < 0$ [5]. The units of $|A|^2$ are expressed in terms of power, typically in milliwatts (mW); since the units of $A(z,T)$ appear in each term, they always cancel and thus $A(z,T)$ need not be assigned units. To express (5.1) in dimensionless form, choose a reference power level P_0 and define $U(z,T) = A(z,T)/\sqrt{P_0}$; here P_0 might be $P_0 = A(0,0)^2 = f(0)^2$ (the "peak" power) or some arbitrary base level such as 1 mW. Next select some reference pulse width T_0, which could be the root mean square width of $f(T)$, the bit period, the full width half-maximum [5], or something else. The corresponding dispersion and nonlinear lengths L_D and L_{NL} are defined by $L_D = T_0^2/|\beta_2|$ and $L_{NL} = 1/(\gamma P_0)$. It is convenient to define dimensionless length and time by $\zeta = z/L_D$ and $t = T/T_0$ [5]. Introducing the constant $N = \sqrt{L_D/L_{NL}}$ and making the appropriate substitutions in (5.1) we obtain the dimensionless nonlinear Schrödinger equation

$$iU_\zeta + \frac{1}{2} U_{tt} + N^2 |U|^2 U = 0, \quad U(0,0) = f(0)/\sqrt{P_0}. \tag{5.2}$$

The N can be removed from (5.2) by making the substitution $u(\zeta,t) = NU(\zeta,t)$, $u(0,t) = q(t)$, bringing (5.2) to the form (1.2). The connection between the physical and dimensionless versions of (1.6) is

$$I = \int_{-\infty}^{\infty} q(t)\,dt = \frac{N}{T_0\sqrt{P_0}} \int_{-\infty}^{\infty} f(T)\,dT, \tag{5.3}$$

in which the reference parameters P_0 and T_0 cancel, since $N^2 = T_0^2 \gamma P_0/|\beta_2|$, leaving

$$I = \int_{-\infty}^{\infty} q(t)\,dt = \sqrt{\frac{\gamma}{|\beta_1|}} \int_{-\infty}^{\infty} f(T)\,dT. \tag{5.4}$$

For single lobe potentials the exact soliton formation criterion is thus $\sqrt{\gamma/|\beta_2|}$ $\int_{-\infty}^{\infty} f(T)\,dT > \pi/2$ by (5.4). Using Theorem 3.3 this extends to higher order solitons using the thresholds $(2n-1)\pi/2$. If $f(T)$ is complex-valued, Corollary 4.3 applies.

Note that (5.4) depends on the physical constants of the fiber and the pulse shape, and not on the particular choices of P_0 and T_0 used in the normalization.

The criterion $\frac{1}{2} < N < \frac{3}{2}$ [5] mentioned above for the hyperbolic secant pulse is only approximate for other shapes. If $f(T)$ is given, $f(0) = \sqrt{P_0}$, and T_0 is the intensity half width at $1/e$ [5], then we can derive specific criteria using (5.3) and (5.4), and detailed knowledge of $\int_{-\infty}^{\infty} f(T)\,dT$. As examples, for Gaussian pulses $f(T) = \sqrt{P_0}\exp\{-\frac{t^2}{2T_0^2}\}$ in (5.2) [5] a fundamental soliton will evolve if $N > \sqrt{\pi/8} \cong 0.6267$, while for a rectangular pulse of height $f(0) = \sqrt{P_0}$ and total width $2T_0$ the corresponding condition is $N > \pi/4 \cong 0.7854$. Analogous threshold levels can be obtained for higher order solitons.

We need to mention some related literature. Kivshar [12] established the $\pi/2$ threshold result for the special case where all EVs corresponding to the potential $\alpha q(t)$ appear on the imaginary axis at $\xi = 0$ as the real parameter α is varied. This approach reduces to analytically solving (1.1) for $\xi = 0$. In general, however, EVs can hop onto the imaginary axis from the upper half plane, without passing through $\xi = 0$ [3]. Kivshar's paper was intended to extend results of Burzlaff [13] which established the bounds in our Theorem 3.3 for the positive "box" and two-sided exponential potentials. Kaup and Scacca [14] also studied potentials made from combining boxes, both positive and negative, a special case of which is the odd potential. They found for odd potentials that EVs are very near the imaginary axis but with definitely nonzero real parts, in contrast to earlier numerical studies which could not confirm that the EVs were not actually on the imaginary axis; see the references in [14]. Theorem 3.4 rules out imaginary EVs for general odd $q(t)$.

Desaix et al. [15] have provided a variational method for approximating ZS EVs. Problems analogous to those discussed in this paper have also been addressed in [16] in connection with the modified KdV equation.

Acknowledgment. We would like to thank a referee for pointing out a paper of Beals and Coifman [17] which treats a more general system than (1.1) and obtains an upper bound on E in (1.4) that rules out EVs. For the particular case of (1.1) the bound in [17] reduces to $E < 1/\sqrt{2}$.

REFERENCES

[1] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, SIAM Stud. Appl. Math 4, SIAM, Philadelphia, 1981.
[2] S. Novikov, S. V. Manikov, L. P. Pitaevskii, and V. E. Zakharov, *Theory of Solitons: The Inverse Scattering Method*, Consultants Bureau, New York, 1984.
[3] M. Klaus and J. K. Shaw, *Influence of pulse shape and frequency chirp on stability of optical solitons*, Opt. Comm., 197 (2001), pp. 491–500.
[4] M. Klaus and J. K. Shaw, *Purely imaginary eigenvalues of Zakharov-Shabat systems*, Phys. Rev. E. (3), 65 (2002), article 036607.
[5] G. P. Agrawal, *Nonlinear Fiber Optics*, 3rd ed., Academic Press, New York, 2001.
[6] V. E. Zakharov and A. B. Shabat, *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media*, Soviet Physics JTEP, 34 (1972), pp. 62–69.
[7] Z. V. Lewis, *Semiclassical solutions of the Zakharov-Shabat scattering problem for phase modulated potentials*, Phys. Lett. A, 112 (1985), pp. 99–103.
[8] J. W. Miles, *An envelope soliton problem*, SIAM J. Appl. Math., 41 (1981), pp. 227–230.
[9] W. Walter, *Ordinary Differential Equations*, Springer-Verlag, New York, 1998.
[10] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vol. 4, Academic Press, New York, 1980.
[11] E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag, New York, 1965.

[12] Y. Kivshar, *On the soliton generation in optical fibers*, J. Phys. A, 22 (1989), pp. 337–340.
[13] J. Burzlaff, *The soliton number of optical soliton bound states for two special families of input pulses*, J. Phys. A, 21 (1988), pp. 561–566.
[14] D. Kaup and L. Scacca, *Generation of* 0-π *pulses from a zero-area pulse in coherent pulse propagation*, J. Opt. Soc. Amer., 70 (1980), pp. 224–230.
[15] M. Desaix, D. Anderson, M. Lisak, and M. Quiroga-Teixeiro, *Variationally obtained approximate eigenvalues of the Zakharov-Shabat scattering problem for real potentials*, Phys. Lett. A, 212 (1996), pp. 332–338.
[16] S. Clarke, R. Grimshaw, P. Miller, E. Pelinovsky, and T. Talipova, *On the generation of solitons and breathers in the modified Korteweg-de Vries equation*, Chaos, 10 (2000), pp. 383–392.
[17] R. Beals and R. R. Coifman, *Scattering and inverse scattering for first order systems*, Comm. Pure Appl. Math., 37 (1984), pp. 39–90.

SIAM J. MATH. ANAL.
Vol. 34, No. 4, pp. 774–804

THE EXISTENCE AND LARGE TIME BEHAVIOR OF SOLUTIONS TO A SYSTEM RELATED TO A PHASE TRANSITION PROBLEM*

HARUMI HATTORI†

Abstract. We discuss the existence and the asymptotic behavior of weak solutions to a hyperbolic-elliptic mixed type system related to a phase transition problem. As in the hyperbolic case, the weak solutions are not unique. To select the admissible solution, we need to impose admissibility criteria. One of the criteria we use is the entropy rate admissibility criterion. The question is how it can be applied. We examine two alternatives, and the result is used to study the existence and large time behavior of solutions to the perturbed Riemann problem.

Key words. phase transition, entropy rate admissibility criterion, entropy condition, hyperbolic-elliptic mixed type, Glimm scheme

AMS subject classifications. 35L65, 35M10, 73B99

PII. S0036141001391378

1. Introduction.

In this paper, we consider the existence and large time behavior of weak solutions to a hyperbolic-elliptic mixed system related to a phase transition problem. The system is given by

$$\begin{aligned} v_t - u_x &= 0, \\ u_t - f(v)_x &= 0, \end{aligned} \tag{1.1}$$

where v, u, and f are strain, velocity, and stress, respectively. We assume that f is a smooth nonmonotone function of v, as depicted in Figure 1.1. The horizontal line for which the areas A and B are equal is called the Maxwell stress, and the strains in the α- and β-phases corresponding to the Maxwell stress are denoted by v_α and v_β, respectively. It is important to note that, if f' is nonnegative, the system is hyperbolic, and f' is negative, then the system is elliptic. In our case, there are two intervals $(0,\alpha]$ and $[\beta,\infty)$, where the system is hyperbolic. They are called the α-phase and β-phase, respectively. In thermodynamics, the states $(0,v_\alpha]$ and $[v_\beta,\infty)$ are said to be stable, $(v_\alpha,\alpha]$ and $[\beta,v_\beta)$ are said to be metastable, and (α,β) is called the spinodal region and is physically unobservable. This is one of the simplest systems capable of explaining a phase transition problem.

One goal is to examine the entropy rate admissibility criterion. As in the hyperbolic systems of conservation laws, the weak solutions for the above mixed type problem are not unique. Therefore, to select a physically relevant solution, the admissibility criteria should be imposed. In section 2, we discuss the main admissibility criteria we use in this paper. They are the entropy condition and the entropy rate admissibility criterion. The entropy rate admissibility criterion, proposed by Dafermos [8], [9] for the hyperbolic systems of conservation laws, roughly says that the rate of (mathematical) entropy production is the smallest for the admissible solution. This criterion is also used in crack dynamics [38], and the effort should be made to extend

*Received by the editors June 25, 2001; accepted for publication (in revised form) July 30, 2002; published electronically February 6, 2003. This work was supported by NSF grant DMS-9704383 and Army DEPSCoR grant DAAD19-02-1-0206.
http://www.siam.org/journals/sima/34-4/39137.html
†Department of Mathematics, West Virginia University, Morgantown, WV 26506-6310 (hhattori@wvu.edu).

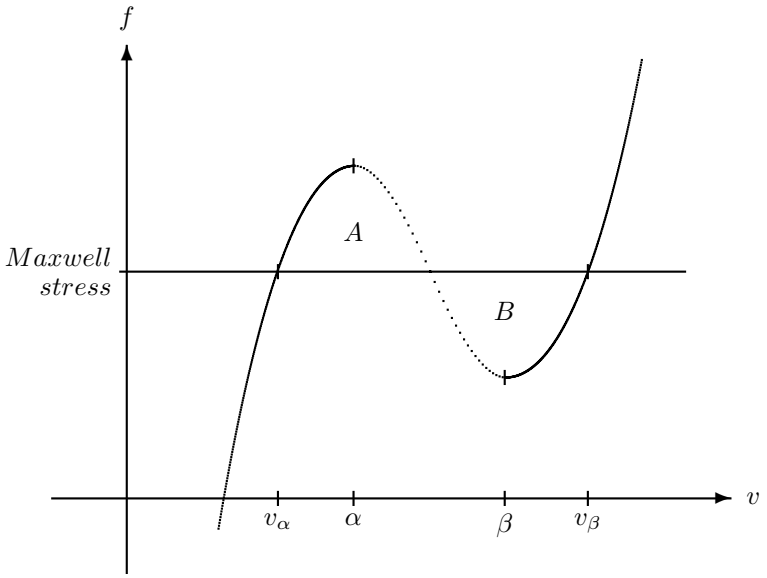


FIG. 1.1.

it to the conservation laws of mixed type. We apply it to the Riemann problem involving the phase boundary. In this case, there are at least two different ways to apply the criterion. Should it be applied to all waves combined or to the phase boundary only? To see the relation, in section 3, we apply the criterion to all of the waves in the Riemann problem to obtain the existence of global weak solutions, and, in section 4, we analyze its consequence to the Riemann problem. We show that, if we apply the criterion in the above fashion to the Riemann problem and take the hyperbolic limit, we obtain the solution to the Riemann problem, where the criterion is applied to the phase boundary only.

Another goal is to discuss the existence and large time behavior of solutions to (1.1) in the space of bounded variations. The initial data are given by

$$U(x,0) = (v,u)(x,0) = (v,u)_{oo}(x) \equiv \begin{cases} U_L = (v_L, u_L), & x \le -M, \\ U_l(x) = (v_l(x), u_l(x)), & -M < x < 0, \\ U_r(x) = (v_r(x), u_r(x)), & 0 < x < M, \\ U_R = (v_R, u_R), & M < x, \end{cases} \tag{1.2}$$

where (v_L, u_L) and (v_R, u_R) are constant states and M is a positive constant. This problem is called the perturbed Riemann problem, and the corresponding Riemann problem is given by

$$(v,u)(x,0) = \begin{cases} U_L = (v_L, u_L), & x < 0, \\ U_R = (v_R, u_R), & 0 < x. \end{cases} \tag{1.3}$$

In what follows, the vector notation U (or V) and the component notation (v,u) are used interchangeably, and the Riemann problem with initial data 1.3 is denoted by (U_L, U_R). We assume the following conditions for the initial data:

1. $U_l(x)$ and $U_r(x)$ are close to U_L and U_R, respectively, in the total variation norm.

2. $v_l(x)$ and v_L are in the α-phase and close to v_α, $v_r(x)$ and v_R are in the β-phase and close to v_β, and $u_{oo}(x)$ is small in total variation.

Therefore, we assume that

$$\begin{aligned}\eta = {} & TV_{x<0}|U_l(x) - U_L| + TV_{x>0}|U_r(x) - U_R| \\ & + |v_L - v_\alpha| + |v_R - v_\beta| + |u_L - u_c| + |u_R - u_c|\end{aligned} \tag{1.4}$$

is small, where u_c is a constant. In particular, we assume that $\eta < 1$. The above initial data will ensure that the speed of the phase boundary is subsonic; namely, its speed is smaller than those of characteristics. We use the Glimm scheme with the admissibility criteria discussed above. A crucial step in the Glimm scheme is to estimate the strengths of outgoing waves for the wave interactions involving the phase boundary. We need to select the outgoing waves so that both the entropy condition and the entropy rate admissibility criterion are satisfied. Concerning the asymptotic behavior, we show that, among other things, the solution approaches the solution of the corresponding Riemann problem modulo shift. The details will be discussed in sections 3 and 5.

It is now a common practice to use the nonmonotone constitutive relation to formulate the conservation laws with phase change. In the inviscid approach, the Riemann problem of system (1.1) was discussed in various literature. James [23] initiated the Riemann problem for this type of problem. Different admissibility criteria were used to select a physically relevant solution. Abeyaratne and Knowles [1], [2] discussed it using the kinetic relation and the initiation criterion. Hattori [17], [18] used the entropy rate admissibility criterion proposed by Dafermos [8], [9] for hyperbolic systems. Shearer [33] considered the problem, assuming that all of the stationary phase boundaries are admissible. Keyfitz [24] discussed the Riemann problem from the point of view of the "hysteresis" approach. Mercier and Piccoli [29] classified the initial data using the kinetic relation. As far as the Cauchy problem is concerned, Le Floch [26] has shown the existence of global solutions for a trilinear system in the space of bounded variations. Asakura [4] considered the nonlinear case. Pego and Serre [31] considered the instability of the Glimm scheme. Colombo and Corli [6] studied the continuous dependence of solutions. Corli and Sablé-Tougeron [7] discussed the sonic phase boundary problem. Another approach is to add the higher spatial derivatives of v and u to smooth out the shock discontinuities and phase boundaries. Slemrod [35], [36] discussed the effects of viscosity and capillarity and proposed the viscosity-capillarity criterion. Shearer [34] considered the issue of nonuniqueness for the Riemann problem using this criterion. Slemrod [37] also discussed the limiting viscosity approach. Fan extended this approach and obtained a series of results [11], [12], [13]. The results of Fan and Slemrod are summarized in [14]. Hattori and Mischaikow [20] considered the soft loading problem with viscosity and capillarity. Hsiao [22], Hoff and Khodja [21], and Pego [30] considered the role of the viscosity.

This paper consists of five sections. In section 2, we describe the admissibility criteria that we use in this paper. They are the entropy condition, the entropy rate admissibility criterion, and the initiation criterion. We then study the Riemann problem with single phase boundary using these criteria. In section 3, we discuss the wave interactions in a diamond and construct the interaction potential. Then we show the existence of weak solutions in the space of bounded variations. In section 4, we revisit the Riemann problem and show that asymptotic states occur immediately in the Riemann problem. We study the large time behavior of weak solutions in section 5.

2. The Riemann problem. In this section, we first describe the waves appearing in the Riemann problem and the admissibility criteria, and then we study the Riemann problem for (1.1) with single phase boundary.

2.1. Waves in the Riemann problem.

1. *Elementary waves.* We call the rarefaction wave and the shock wave the elementary waves. The 1-rarefaction curve $R_1^r(U_o)$ and the 1-shock curve $S_1^r(U_o)$ through U_o are the set of U connected to U_o on the right by the respective waves. They satisfy the following relations:

$$\text{rarefaction curve:} \quad u = u_o + \int_{v_o}^{v} \lambda(w)dw, \qquad \begin{cases} v \le v_o & \text{if } f \text{ is convex,} \\ v \ge v_o & \text{if } f \text{ is concave,} \end{cases}$$

$$\text{shock curve:} \quad u = u_o - \sigma_b(v_o, v)(v - v_o), \qquad \begin{cases} v \ge v_o & \text{if } f \text{ is convex,} \\ v \le v_o & \text{if } f \text{ is concave,} \end{cases}$$

where $\lambda(w) = \sqrt{f'(w)}$ and $\sigma_b(v_o, v) = -\sqrt{\frac{f(v)-f(v_o)}{v-v_o}}$. The combined wave curve is denoted by $T_1^r(U_o)$. The 2-rarefaction, 2-shock, and combined curves $R_2^r(U_o)$, $S_2^r(U_o)$, and $T_2^r(U_o)$ are defined in a similar manner:

$$\text{rarefaction curve:} \quad u = u_o - \int_{v_o}^{v} \lambda(w)dw, \qquad \begin{cases} v \ge v_o & \text{if } f \text{ is convex,} \\ v \le v_o & \text{if } f \text{ is concave,} \end{cases}$$

$$\text{shock curve:} \quad u = u_o - \sigma_f(v_o, v)(v - v_o), \qquad \begin{cases} v \le v_o & \text{if } f \text{ is convex,} \\ v \ge v_o & \text{if } f \text{ is concave,} \end{cases}$$

where $\sigma_f(v_o, v) = \sqrt{\frac{f(v)-f(v_o)}{v-v_o}}$. We define $R_1^l(U_o)$, $S_1^l(U_o)$, $T_1^l(U_o)$, $R_2^l(U_o)$, $S_2^l(U_o)$, and $T_2^l(U_o)$ as the sets of U connected to U_o on the left by the corresponding waves. If the above inequalities are reversed, we obtain the corresponding relations. We measure the wave strength of the elementary waves by $\pm|\lambda(v) - \lambda(v_o)|$, where the plus sign is for the rarefaction waves and the minus sign is for the shock waves. A collection of the 1-waves or 2-waves is called a family of waves.

2. *Phase boundary.* A phase boundary is the line of discontinuity in the xt-plane across which the phase changes. It satisfies the Rankine–Hugoniot condition. The phase boundary curve $P^r(U_o)$ (or $P^l(U_o)$) is the set of U connected to U_o on the right (or left) by the phase boundary and satisfies the relation

$$u = u_o - \sigma_p(v_o, v)(v - v_o),$$

where $\sigma_p(v_o, v) = \pm\sqrt{\frac{f(v)-f(v_o)}{v-v_o}}$ and v_o and v are in the different phases. We measure the wave strength of the phase boundary by $|v - v_o|$. This gives the variation of phase boundary and is equivalent to the wave strengths for shocks and rarefaction waves.

2.2. Admissibility criteria. The weak solutions for (1.1) are not unique, and we use admissibility criteria to choose a physically relevant solution. There are three criteria that we use in this paper. They are the entropy condition, the entropy rate admissibility criterion, and the initiation criterion. The entropy (physically, the energy) for (1.1) is given by

$$H = \frac{1}{2}u^2 + \int f(v)dv.$$

The rate of decay of the total energy is given by

$$D_+H = \sum_{\text{jump discontinuties}} \sigma(v_-, v_+)A(v_-, v_+), \tag{2.1}$$

where $\sigma(v_-, v_+)$ is the speed of the jump discontinuity and

$$A(v_-, v_+) = \left[\frac{1}{2}(f(v_-) + f(v_+))(v_+ - v_-) - \int_{v_-}^{v_+} f(w)dw\right].$$

Here v_- and v_+ are the values of v on the left and right of a jump discontinuity. We denote

$$E(v_-, v_+) = \sigma(v_-, v_+)A(v_-, v_+).$$

The entropy condition requires that $E(v_-, v_+) \leq 0$ is satisfied across a shock or a phase boundary.

The entropy rate admissibility criterion is the criterion that was proposed by Dafermos [8], [9]. This criterion roughly says that the rate of entropy production is the smallest for the admissible solution. Specifically, this criterion postulates that the solution is admissible if it solves (1.1) and minimizes (2.1).

The initiation criterion has been used in [1], [4], and [26]. This criterion imposes that no new phase occurs from any point except when no solution exists without the creation of a new phase. This ensures that spontaneous initiation of a new phase cannot occur from two nearby initial states in the same phase.

As discussed in the introduction, there are at least two ways to apply the entropy rate admissibility criterion in the Riemann problem. Before the separation ($t = 0_-$) of the waves, they influence each other. Therefore, it can be applied to all jump discontinuities in the Riemann problem. In other words, we compute

$$\min \sum_i E(v_{i-}, v_{i+}), \tag{2.2}$$

where the index i runs for all of the jump discontinuities in the Riemann problem and v_{i-} and v_{i+} are the values of v at the ith discontinuity. If the criterion is applied this way, we say that the entropy rate admissibility criterion is applied before the separation. Also, it can be applied to each phase boundary after the separation. In this case, if we take the adjacent states to a phase boundary as the initial data, the solution to the Riemann problem produces that phase boundary only. In other words, the phase boundary is stable in the sense that by itself it is the admissible solution of the Riemann problem if the adjacent states are the initial data. Note that the shocks are stable in this sense. Once the waves in the Riemann problem are separated, it may not be reasonable to apply the criterion to all of the jump discontinuities. If the criterion is applied this way, we say that the entropy rate admissibility criterion is applied after the separation. In what follows, we apply the entropy rate admissibility criterion to all of the jump discontinuities in the Riemann problem where the phase boundary is involved and obtain the existence and asymptotic behavior of weak solutions. Then we discuss the relation between the two.

2.3. The Riemann problem. In this subsection, we discuss the Riemann problem, where the entropy rate admissibility criterion is applied to all discontinuities. The initial data are given by (U_l, U_r), where v_l and v_r are given in the different phases.

We assume that v_l is in the α-phase and v_r is in the β-phase. We require that v_l and v_r are close to v_α and v_β, respectively, and that u_l and u_r are close. We look for a self-similar solution in which the constant states U_l, U_1, U_2, and U_r are separated by the 1-wave, the phase boundary, and the 2-wave. This is based on the fact that, if there are three or more phase boundaries in the solution of the Riemann problem, at least one of them violates the entropy condition [1], [19]. From the entropy condition, we impose $\sigma_p A(v_1, v_2) \le 0$ across the phase boundary. This condition is necessary because it has been shown in [19] that, unlike the hyperbolic conservation laws, the entropy rate admissibility criterion is not consistent with the entropy condition. We also require that the speed of the phase boundary in absolute value be less than or equal to that of the 1- and 2-waves. If this condition is violated, we may have a geometrically inconsistent solution. The above considerations motivate the following minimization problem:

$$\min \{E_b + E_p + E_f\} \tag{2.3}$$

subject to the entropy condition

$$\sigma_p A(v_1, v_2) \le 0, \tag{2.4}$$

the characteristic condition

$$\sigma_b \text{ or } -\lambda_1 \le \sigma_p \le \sigma_f \text{ or } \lambda_2, \tag{2.5}$$

and

$$\begin{aligned} u_l + \left\{ \begin{array}{l} -\sigma_b(v_1 - v_l),\ v_l > v_1 \\ \int_{v_l}^{v_1} \lambda_1(w)dw,\ v_l \le v_1 \end{array} \right\} \\ - \sigma_p(v_2 - v_1) - \left\{ \begin{array}{l} \sigma_f(v_r - v_2),\ v_r < v_2 \\ \int_{v_2}^{v_r} \lambda_2(w)dw,\ v_r \ge v_2 \end{array} \right\} = u_r, \end{aligned} \tag{2.6}$$

where

$$E_p = \sigma_p A(v_1, v_2),$$

$$E_b = \left\{ \begin{array}{ll} \sigma_b A(v_l, v_1), & v_l > v_1, \\ 0, & v_l \le v_1, \end{array} \right.$$

$$E_f = \left\{ \begin{array}{ll} \sigma_f A(v_2, v_r), & v_r < v_2, \\ 0, & v_r \ge v_2. \end{array} \right.$$

The condition (2.6) states the way in which U_l, U_1, U_2, and U_r are connected. We choose v_1 as the independent variable and derive the differential equations governing v_2 and the entropy rate. The admissible solution is the solution to the Riemann problem (2.3)–(2.6). We say that a solution is feasible if it satisfies (2.4) and (2.5). In [32], this type of problem was discussed in the case where there are no shock waves. The region of U_1 where (2.4) and (2.5) are satisfied for a given 2-wave curve $T^l(U_r)$ is called the feasible region. The v-coordinate of the intersection between the 1-wave curve and the feasible region gives the interval of v_1 in which the solution satisfies (2.4) and (2.5). Corresponding to the equality signs in (2.4) and (2.5), we define the following curves.

DEFINITION 2.1. *The stationary phase boundary curve (SC): This is the set of U_1 satisfying $\sigma_p = 0$ and is given by*

$$f(v_1) = f(v_2), \quad u_1 = u_2 \tag{2.7}$$

as U_2 moves along the 2-wave curve.

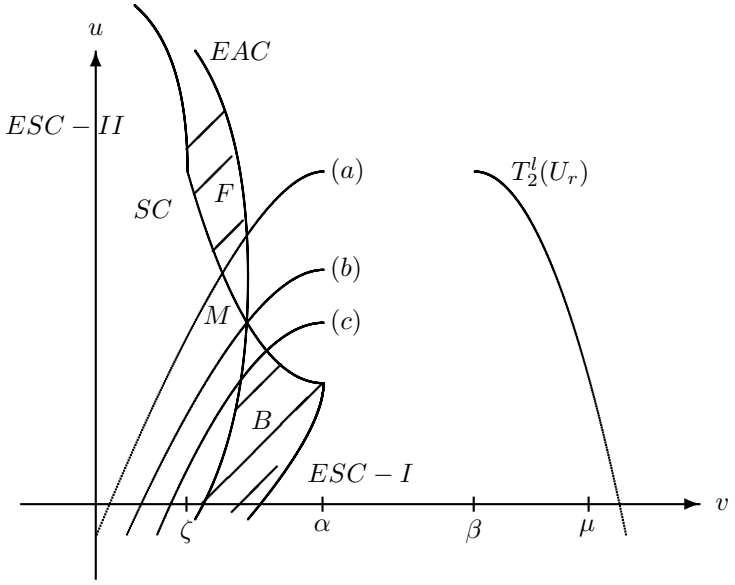


FIG. 2.1.

DEFINITION 2.2. *The equal area curve (EAC): This is the set of* U_1 *satisfying*

$$A(v_1, v_2) = 0, \quad u_1 = u_2 + \sigma_p(v_2 - v_1) \tag{2.8}$$

as U_2 *moves along the* 2*-wave curve.*

DEFINITION 2.3. *The equal speed curve-I (ESC-I): This is the set of* U_1 *satisfying*

$$\sigma_p = \begin{cases} -\lambda_1, & v_1 \geq v_l, \\ \sigma_b, & v_1 < v_l, \end{cases} \tag{2.9}$$

$$u_1 = u_2 + \sigma_p(v_2 - v_1) \tag{2.10}$$

as U_2 *moves along the* 2*-wave curve. This curve starts from* (α, u_1)*, where*

$$u_1 = u_r + \begin{cases} \sigma_f(v_r - \mu), & v_r < \mu, \\ \int_\mu^{v_r} \lambda_2(w)dw, & v_r \geq \mu. \end{cases}$$

If $v_1 < v_l$*, the line segment in the strain-stress plane joining* $(v_2, f(v_2))$ *and* $(v_1, f(v_1))$ *passes through* $(v_l, f(v_l))$*.*

DEFINITION 2.4. *The equal speed curve-II (ESC-II): This is the set of* U_1 *satisfying*

$$\sigma_p = \begin{cases} \lambda_2, & v_2 \leq v_r, \\ \sigma_f, & v_2 > v_r, \end{cases} \tag{2.11}$$

$$u_1 = u_2 + \sigma_p(v_2 - v_1) \tag{2.12}$$

as U_2 *moves along the* 2*-wave curve. This curve starts from* (γ, u_1)*, where* $u_1 = u_r + \int_\beta^{v_r} \lambda_2(w)dw$*. If* $v_r < v_2$*, the line segment in the strain-stress plane joining* $(v_2, f(v_2))$ *and* $(v_1, f(v_1))$ *passes through* $(v_r, f(v_r))$*.*

The curves satisfying the above definitions for the 2-wave curve $T_2^l(U_r)$ are depicted in Figure 2.1. The feasible regions are shaded regions. Depending on how the 1-wave curve $T_1^r(U_l)$ intersects with the shaded regions, we obtain three cases.

(a) The 1-wave curve intersects with the region F in Figure 2.1.

(b) The 1-wave curve goes through the point M in Figure 2.1.

(c) The 1-wave curve intersects with the region B in Figure 2.1.

The v-coordinate of M is v_α, and F (or B) stands for the fact that the phase boundary moves forward (or backward) if the 1-wave curve intersects these regions. For example, in case (a), the solutions are feasible if U_1 is on the intersection of the curve (a) with the shaded region, and we seek a solution to (2.3) and (2.6) for the values of v_1 on the interval where the 1-wave curve (a) intersects with F.

The following theorem shows the existence of solutions satisfying (2.3) and (2.6).

THEOREM 2.5 (Hattori [19]). *There exists an absolute minimum for the Riemann problem* (2.3)–(2.6). *Furthermore, there exists a neighborhood of* $v_l = v_\alpha$, $v_r = v_\beta$, $u_l = u_c$, *and* $u_r = u_c$ *such that the problem has a unique admissible solution.*

3. Existence of weak solutions. This section consists of three subsections. First, we briefly describe the Glimm scheme and its two main ingredients. Then, in subsections 3.2 and 3.3, we discuss the details of the ingredients leading to the proof of the existence of weak solutions.

3.1. Glimm scheme. We choose a sequence $\omega = \{\omega_n\}$ of random and equidistributed numbers in $(-1, 1)$ and define

$$N_{m,n} = ((m + \omega_n)\triangle x, n\triangle t), \quad n \geq 0,$$

as the sample points, where m and n are integers satisfying $m + n =$ odd. Here $\triangle x$ and $\triangle t$ are positive numbers satisfying the Lax–Friedrichs condition

$$\frac{\triangle x}{\triangle t} = r > \max_{v \in V_{v_\alpha} \cup V_{v_\beta}} \lambda(v),$$

where V_{v_α} and V_{v_β} are closed neighborhoods of v_α and v_β, respectively, relevant to our discussion. The upper xt-plane is divided by diamond-shaped domains $\triangle_{m,n}$ with vertices $N_{m,n+1}$, $N_{m-1,n}$, $N_{m,n-1}$, $N_{m+1,n}$. An I-curve is a space-like piecewise-linear curve composed of line segments joining vertices. We define J_0 to be the I-curve consisting of the line segments joining $N_{m-1,0}$, $N_{m,1}$, $N_{m+1,0}$ with m even. It is possible to have a partial ordering of the I-curves. If J_2 lies in the future of J_1, it is denoted by $J_1 \leq J_2$. We say that J_1 and J_2 are consecutive if $J_1 \leq J_2$ and there is only one diamond between them.

The approximate solution $U_{\triangle x,\omega}$ is defined as follows. We choose

$$U_{\triangle x,\omega}(N_{m,o}) = U_{oo}((m + \omega_o)\triangle x).$$

Assuming that $U_{\triangle x,\omega}$ is defined at $N_{m-1,n-1}$ and $N_{m+1,n-1}$, we solve the Riemann problem of (1.1) with the initial data

$$U(x, (n-1)\triangle t) = \begin{cases} U_{\triangle x,\omega}(N_{m-1,n-1}), & (m-1)\triangle x \leq x < m\triangle x, \\ U_{\triangle x,\omega}(N_{m+1,n-1}), & m\triangle x < x \leq (m+1)\triangle x, \end{cases}$$

and the solution is denoted as $U(x, t)$. We then define the solution on the next time level by

$$U_{\triangle x,\omega}(N_{m,n+1}) = U(N_{m,n+1}),$$

$$U_{\triangle x,\omega}(x,t) = U(x,t), \quad (m-1)\triangle x \le x \le (m+1)\triangle x, \quad n\triangle t < t < (n+1)\triangle t.$$

The wave strengths of the solution to the Riemann problem are denoted as

$$a = (a_1, a_2) \quad \text{if there is no phase boundary, and}$$
$$a = (a_1, P, a_2) \quad \text{if there is a phase boundary,}$$

where the subscripts 1 and 2 denote the 1- and 2-waves, respectively. The outgoing waves c are the waves in the solution of the Riemann problem in $\triangle_{m,n}$. The incoming waves a and b to $\triangle_{m,n}$ are the outgoing waves from $\triangle_{m-1,n-1}$ and $\triangle_{m+1,n-1}$, respectively, and entering $\triangle_{m,n}$. If $a = (a_1, a_2)$ and $b = (b_1, b_2)$ cross J and a is lying to the left of b, we say that a_i and b_j are approaching provided that either (i) $i > j$ or (ii) $i = j$ and at least one of them is a shock.

Let $P(J)$ be the strength of phase boundary crossing J, and let $W(J)$ be the collection of the elementary waves crossing J. We define

$$L(J) = \sum_{a \in W(J)} |a| .$$

Then $L(J) + P(J)$ measures the total variation of $U_{\triangle x,\omega}(x,t)$ on J. The Glimm scheme consists of two steps. The first step is to estimate the strengths of outgoing waves in terms of incoming waves in a diamond. The second step is to show that the total variation of the solutions is bounded. Since the total variation is not necessarily monotonically decreasing, we define the interaction potential $Q(J)$ decreasing in J. With this Q we show that the inequality

$$L(J_2) + P(J_2) + Q(J_2) \le L(J_1) + P(J_1) + Q(J_1)$$

holds for $J_1 < J_2$. The details of these steps are discussed in the next two subsections.

3.2. Local wave interaction. We consider the wave interaction in a diamond $\triangle_{m,n}$ and estimate the strengths of outgoing waves in terms of those of incoming waves. There are two cases depending on whether the phase boundary enters the diamond. First, consider the case where the phase boundary does not enter the diamond. This case is treated in the same way as in the previous literature. Let (a_1, a_2) and (b_1, b_2) be the strengths of incoming waves from $\triangle_{m-1,n-1}$ and $\triangle_{m+1,n-1}$, respectively, and let (c_1, c_2) be the strengths of outgoing waves.

LEMMA 3.1 (Glimm [15] and Liu [27]). *Suppose that the phase boundary P does not enter the diamond. Then we have*

$$c_i = a_i + b_i + Q(\triangle_{m,n}), \quad i = 1, 2, \tag{3.1}$$

where $Q(\triangle_{m,n})$ is defined as

$$Q(\triangle_{m,n}) = Q_s(\triangle_{m,n}) + Q_d(\triangle_{m,n}), \tag{3.2}$$

where

$$Q_s(\triangle_{m,n}) = \begin{cases} 0, & a_i \ge 0,\ b_i \ge 0, \\ |a_i|^3, & |a_i| \le |b_i|,\ a_i < 0,\ b_i \ge 0, \\ |a_i||b_i|^2, & |a_i| \le |b_i|,\ b_i < 0, \\ |a_i|^2|b_i|, & |a_i| \ge |b_i|,\ a_i < 0, \\ |b_i|^3, & |a_i| \ge |b_i|,\ a_i \ge 0,\ b_i < 0, \end{cases} \tag{3.3}$$

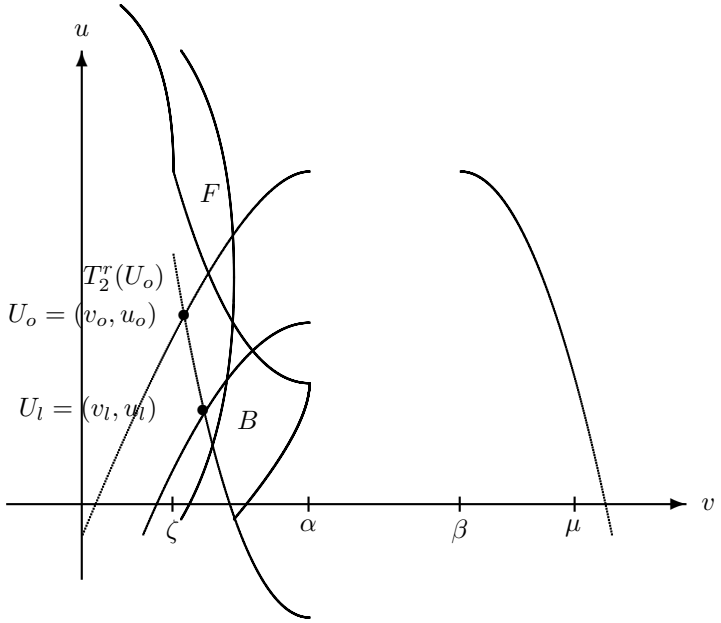


FIG. 3.1.

or

$$= \begin{cases} 0, & a_i \geq 0,\ b_i \geq 0, \\ |a_i||b_i|(|a_i| + |b_i|), & \textit{otherwise,} \end{cases} \tag{3.4}$$

$$Q_d(\triangle_{m,n}) = |a_2 b_1|.$$

Next, we consider the case where the phase boundary enters the diamond $\triangle_{m,n}$. Without loss of generality, we assume that the phase boundary enters $\triangle_{m,n}$ from $\triangle_{m+1,n-1}$. The wave strengths of the incoming waves from $\triangle_{m-1,n-1}$ and $\triangle_{m+1,n-1}$ are denoted by (a_1, a_2) and (b_1, P_i, b_2), respectively, and (c_1, P_o, c_2) for the outgoing waves. There are two subcases to consider. One subcase is that the 2-wave from $\triangle_{m+1,n-1}$ enters the diamond, or it is a rarefaction wave. Another subcase is that the 2-wave from $\triangle_{m+1,n-1}$ does not enter the diamond, and it is a shock. The estimates are given in Lemmas 3.2 and 3.4, respectively. Note that, in the first subcase, the incoming waves from $\triangle_{m+1,n-1}$ are the admissible solution to the Riemann problem (2.3)–(2.6), but, in the second subcase they are not. This makes the estimate more complicated.

LEMMA 3.2. *Suppose that all of the outgoing waves from the diamond $\triangle_{m+1,n-1}$ enter $\triangle_{m,n}$, or the 2-wave from $\triangle_{m+1,n-1}$ which (or a portion of which) does not enter $\triangle_{m,n}$ is a rarefaction wave. Then the outgoing waves from $\triangle_{m,n}$ satisfy the following estimates:*

$$\begin{aligned} c_1 &= b_1 + O(1)|a_2|, \\ c_2 &= b_2 + O(1)|a_2|, \\ P_o &= P_i + O(1)|a_2|. \end{aligned} \tag{3.5}$$

Proof. Note that the wave entering $\triangle_{m,n}$ from $\triangle_{m-1,n-1}$ is a_2 only. Denote the constant states of the incoming waves by U_o, U_l, U_1, U_2, and U_r. The wave

a_2 is between U_o and U_l. As we change U_l along the 2-wave curve through U_o, the minimum point U_1 moves along the composite curve consisting of the SC, the EAC, or the curve lying between the SC and the EAC; see Figures 2.1 and 3.1. In what follows, we prove the case in which all of the elementary waves are rarefaction waves. The other cases are proved similarly. If U_1 is on the SC, we have

$$u_o - \int_{v_o}^{v_l} \lambda(w)dw + \int_{v_l}^{v_1} \lambda(w)dw = u_1 = u_2 = u_r + \int_{v_2}^{v_r} \lambda(w)dw,$$

$$f_1 = f_2.$$

If U_1 is on the EAC, we have

$$u_o - \int_{v_o}^{v_l} \lambda(w)dw + \int_{v_l}^{v_1} \lambda(w)dw = u_1 = \sigma_p(v_2 - v_1) + u_r + \int_{v_2}^{v_r} \lambda(w)dw,$$

$$\frac{1}{2}(f(v_1) + f(v_2))(v_2 - v_2) - \int_{v_1}^{v_2} f(w)dw = 0.$$

Also, if U_1 is between the SC and the EAC, we see that

$$u_o - \int_{v_o}^{v_l} \lambda(w)dw + \int_{v_l}^{v_1} \lambda(w)dw = u_1 = \sigma_p(v_2 - v_1) + u_r + \int_{v_2}^{v_r} \lambda(w)dw,$$

$$\frac{dE}{dv_1} = \frac{1}{4\sigma_p}\left\{\frac{dv_2}{dv_1}(\lambda_2^2 - \sigma_p^2)A_{21} - (\lambda_1^2 - \sigma_p^2)A_{12}\right\} = 0.$$

In each case, the mapping from U_l to U_1 is differentiable. Since it can be shown that $\frac{du_1}{dv_l} \le 0$ in each case, U_1 moves monotonically along a Lipschitz curve as we move U_l along the 2-wave curve through U_o. Therefore, we have (3.5). □

To prepare for the second subcase, we need the following lemma.

LEMMA 3.3. *Suppose that $a = 0$ and $b = (b_1, P_i, \bar{b}_2)$, where $\bar{b}_2$ from $\triangle_{m+1,n-1}$ is a shock wave, and it does not enter $\triangle_{m,n}$. If we solve the Riemann problem* (2.3)–(2.6) *with the initial data U_l and U_2, the wave strengths of the outgoing waves from $\triangle_{m,n}$ satisfy*

$$|c_1 - b_1| = O(1)|\bar{b}_2|^2, \quad |P_o - P_i| = O(1)|\bar{b}_2|^2, \quad |c_2| = O(1)|\bar{b}_2|^2. \tag{3.6}$$

Suppose that $a = (\bar{a}_1, P_i, a_2)$ and $b = 0$, where $\bar{a}_1$ from $\triangle_{m-1,n-1}$ is a shock wave, and it does not enter $\triangle_{m,n}$. If we solve the Riemann problem (2.3) *and* (2.6) *with the initial data U_1 and U_r, the wave strengths of the outgoing waves from $\triangle_{m,n}$ satisfy*

$$|c_1| = O(1)|\bar{a}_1|^2, \quad |P_o - P_i| = O(1)|\bar{a}_1|^2, \quad |c_2 - a_2| = O(1)|\bar{a}_1|^2. \tag{3.7}$$

Proof. We prove (3.7) so that we can use v_1 as the independent variable. The case (3.6) is proved similarly. Assume that the Riemann problem with the initial data U_1 and U_r has U_1' and U_2' as the intermediate constant states.

Depending on the location of U_1 relative to the feasible region, we have three cases, as in the previous lemma. We prove the case in which U_1 is on the EAC. The cases in which U_1 is on the SC and between the EAC and the SC can be shown in a similar manner. Since U_l is connected to U_1 by a 1-shock wave, $v_l > v_1$ and $u_l > u_1$ must hold. We treat the case in which the 2-wave is a rarefaction wave and the 1-wave

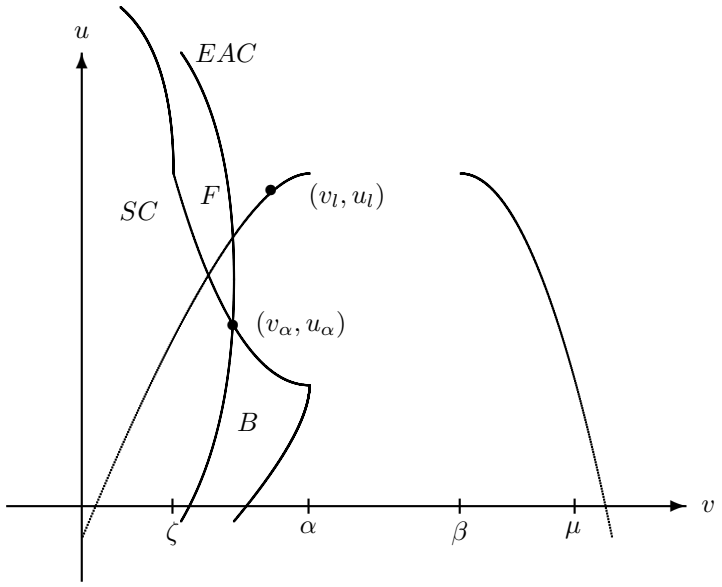


FIG. 3.2.

curve from U_l goes through U_α or above; see Figure 3.2. In this case, we observe the forward phase boundary. The other case is treated in a similar manner. In this case,

$$\frac{dE_b}{dv_1} + \frac{dE_p}{dv_1} + \frac{dE_f}{dv_1}\bigg|_{v_1=v_1} < 0 \quad \text{and} \quad \frac{dE_b}{dv_1}\bigg|_{v_1=v_1} > 0.$$

This implies that

$$A_1 = \frac{dE_p}{dv_1} + \frac{dE_f}{dv_1}\bigg|_{v_1=v_1} < 0.$$

Compare this with

$$A_1' = \frac{dE_p}{dv_1'} + \frac{dE_f}{dv_1'}\bigg|_{v_1'=v_1}.$$

From

$$A_1 = \frac{\partial E_p}{\partial v_1} + \frac{\partial E_p}{\partial v_2}\frac{dv_2}{dv_1} + \frac{\partial E_f}{\partial v_2}\frac{dv_2}{dv_1}\bigg|_{v_1=v_1},$$

$$A_1' = \frac{\partial E_p}{\partial v_1'} + \frac{\partial E_p}{\partial v_2'}\frac{dv_2'}{dv_1'} + \frac{\partial E_f}{\partial v_2'}\frac{dv_2'}{dv_1'}\bigg|_{v_1'=v_1},$$

$$\frac{dv_2}{dv_1}\bigg|_{v_1=v_1} = \frac{(\lambda_1^2 + \sigma_b\sigma_p)(\sigma_b - \sigma_p)}{\sigma_b(\lambda_2 - \sigma_p)^2},$$
$$\frac{dv_2'}{dv_1'}\bigg|_{v_1'=v_1} = \frac{(\lambda_1 - \sigma_p)^2}{(\lambda_2 - \sigma_p)^2},$$

we see that

$$A_1' - A_1 = \left\{ \frac{\partial E_p}{\partial v_2} + \frac{\partial E_f}{\partial v_2} \right\} \frac{\sigma_p(\lambda_1 + \sigma_b)^2}{\sigma_b(\lambda_2 - \sigma_p)^2} = O(1)|\sigma_p||v_1 - v_l|^2.$$

This shows that, if $A_1' < 0$, the configuration does not change, and if $A_1' > 0$,

$$0 < A_1' = A_1 + O(1)|\sigma_p||v_1 - v_l|^2 \le O(1)|\sigma_p||v_1 - v_l|^2.$$

Since $\frac{d^2 E_p}{dv_1'^2} = O(1)$ near the Maxwell construction, we see that

$$|v_1' - v_1| = O(1)|v_1 - v_l|^2.$$

From this we obtain (3.7). □

LEMMA 3.4. *Suppose the 2-wave from $\triangle_{m+1,n-1}$ is a shock and it does not enter $\triangle_{m,n}$. Denote its wave strength by $\bar{b}_2$. Then the outgoing waves from $\triangle_{m,n}$ satisfy the following estimates:*

$$\begin{aligned} c_1 &= b_1 + O(1)|a_2| + O(1)|\bar{b}_2|^2, \\ c_2 &= O(1)|a_2| + O(1)|\bar{b}_2|^2, \\ P_o &= P_i + O(1)|a_2| + O(1)|\bar{b}_2|^2. \end{aligned}$$

Proof. From Lemma 3.3, we see that the wave strengths of the Riemann problem with initial data (v_l, u_l) and (v_3, u_3) corresponding to the waves of the right family $b = \{b_1, P_e, 0\}$ are given by $\{b_1 + O(1)|\bar{b}_2|^2, P_e + O(1)|\bar{b}_2|^2, O(1)|\bar{b}_2|^2\}$. Now we use Lemma 3.2 to obtain the strengths of the outgoing waves. Since the admissible solution is unique, this completes the proof. □

3.3. Existence of global solutions. Let $W(J)$ be the collection of the elementary waves crossing J, and let $W_a(J)$ be the subset of $W(J)$ approaching the phase boundary P. Also, let $W^-(J)$ be the collection of shocks crossing J. We define

$$A(J) = \sum_{a \in W_a(J)} |a|. \tag{3.8}$$

Also, if the phase boundary P crosses J, we define $\triangle_J$ to be the diamond from which P leaves.

Now we define the interaction potential Q as

$$Q(J) = Q_s(J) + N_1\{Q_a(J) + Q_d(J)\} + N_1^2 Q_{aa}(J) + N_2 Q_{sp}(\triangle_J), \tag{3.9}$$

where the quantities on the right-hand side are given by

$$\begin{aligned} Q_s(J) = {} & \textstyle\sum\{|a|^3 : a \in W^-(J)\} \\ & + 4\textstyle\sum\{|bd||b+d| : b,\ d \in W^-(J),\ b \ne d, \text{ and of the same family}\} \\ & + 8\textstyle\sum\{|bde| : b,\ d,\ e \in W^-(J), \text{ distinct, and of the same family}\} \end{aligned} \tag{3.10}$$

or

$$\begin{aligned} Q_s(J) = {} & \textstyle\sum\{|bd||b+d| : b,\ d \text{ are any } j\text{-waves in } W(J) \\ & \text{and not both are } j\text{-rarefaction waves}\}, \end{aligned} \tag{3.11}$$

$$
\begin{aligned}
Q_a(J) &= \sum\{|ab| : a \in W_a(J),\ b \in W(J), \text{ and } a \neq b\},\\
Q_d(J) &= \sum\{|ab| : a,\ b \in W(J) \text{ and } a \text{ is any } i\text{-wave}\\
&\qquad \text{lying toward the left of a } j\text{-wave } b \text{ with } i > j\},\\
Q_{aa}(J) &= \tfrac{1}{2}\sum\{|a|^2 : a \in W_a(J)\} + \sum\{|ab| : a,\ b \in W_a(J) \text{ and } a \neq b\},\\
Q_{sp}(\triangle_J) &= \sum\{|a|^2 : a \text{ is an outgoing wave from } \triangle_J\}.
\end{aligned}
$$

Note that $Q_s(J)$ measures the interaction potential of the waves of the same family, $Q_d(J)$ measures the interaction potential of the waves of the different families, $Q_a(J)$ and $Q_{aa}(J)$ are related to the approaching waves, and $Q_{sp}(\triangle_J)$ is related to the elementary waves leaving from $\triangle_J$. The constants N_1 and N_2 will be determined so that

$$
Q(\triangle_{m,n}) \leq 2(Q(J) - Q(J')) \tag{3.12}
$$

is satisfied, where J and J' are consecutive I-curves with $J \leq J'$, $\triangle_{m,n}$ is the diamond between them, and $Q(\triangle_{m,n})$ is the measure of the wave interactions in $\triangle_{m,n}$. The specific forms of $Q(\triangle_{m,n})$ are given in Lemmas 3.5, 3.6, and 3.7, where the proofs of (3.12) are carried out. Also, let K be a bound on the $O(1)$ coefficients (or a finite multiple of them) appearing in Lemmas 3.1, 3.2, and 3.4.

LEMMA 3.5. *If the phase boundary does not enter $\triangle_{m,n}$, we have*

$$
Q(J') - Q(J) \leq \{-1 + (K + KN_1 + KN_1^2)L(J)\}Q(\triangle_{m,n}), \tag{3.13}
$$

where $Q(\triangle_{m,n})$ is defined in (3.2).

Since $\triangle_J = \triangle_{J'}$ in this case, the proof is reduced to Asakura [3] and Chern [5]. Hence it is omitted. The next two lemmas are similar in proof. Therefore, we prove only Lemma 3.7.

LEMMA 3.6. *If the phase boundary enters $\triangle_{m,n}$ and the 2-wave from $\triangle_J$ $(= \triangle_{m+1,n-1})$ enters $\triangle_{m,n}$ or it is a rarefaction wave, then we have*

$$
Q(J') - Q(J) \leq -Q(\triangle_{m,n}), \tag{3.14}
$$

where $Q(\triangle_{m,n}) = |a_2|\,|b_1| + |a_2|\,|b_2|$.

LEMMA 3.7. *If the phase boundary enters $\triangle_{m,n}$ and the 2-wave with strength $\bar{b}_2$ from $\triangle_J$ $(= \triangle_{m+1,n-1})$ is a shock and it does not enter $\triangle_{m,n}$, then we have*

$$
Q(J') - Q(J) \leq -Q(\triangle_{m,n}), \tag{3.15}
$$

where $Q(\triangle_{m,n}) = |a_2|\,|b_1| + |\bar{b}_2|^2$.

Proof. Let the strengths of the incoming waves from $\triangle_J$ be $b = \{b_1, P_i, 0\}$. Let $A(\triangle_{m,n}) = |a_2|$ be the strength of the approaching wave entering $\triangle_{m,n}$. We prove the case in which both c_i are shocks.

$$
\begin{aligned}
Q_s(J') = {} & \left|b_1 + O(1)|a_2| + O(1)|\bar{b}_2|^2\right|^3\\
& +4 \sum_{d_1 \neq c_1} \left|d(b_1 + O(1)|a_2| + O(1)|\bar{b}_2|^2)\right| \left|d + b_1 + O(1)|a_2| + O(1)|\bar{b}_2|^2\right|\\
& +8 \sum_{d_1, e_1 \neq c_1} \left|de(b_1 + O(1)|a_2| + O(1)|\bar{b}_2|^2)\right|\\
& + \left|O(1)|a_2| + O(1)|\bar{b}_2|^2\right|^3
\end{aligned}
$$

$$
\begin{aligned}
&+4\sum_{d_2\neq c_2}\left|d(O(1)|a_2|+O(1)|\bar{b}_2|^2)\right|\left|d+O(1)|a_2|+O(1)|\bar{b}_2|^2\right|\\
&+8\sum_{d_2,e_2\neq c_2}\left|de(O(1)|a_2|+O(1)|\bar{b}_2|^2)\right|+O.T.\\
&\leq Q_s(J)+K\sum_{e\in W(J)}|e|A(\triangle_{m,n})+K\sum_{e\in W(J)}|e||\bar{b}_2|^2+O.T.,
\end{aligned}
$$

where $O.T.$ represents the potentials not involving the waves in $\triangle_{m,n}$ and canceling out when we estimate $Q_s(J')-Q_s(J)$. Since Q_s is the third order term, we can get the above type estimates easily in the other cases. In what follows, we omit the $O.T.$'s. The estimates of Q_a are given by

$$
\begin{aligned}
Q_a(J')\leq&\sum_{e\in W_a(J_l\cup J_r)}|eb_1|+\sum_{e\in W_a(J_l\cup J_r)}|eb_2|\\
&+K\sum_{e\in W_a(J_l\cup J_r)}|e|A(\triangle_{m,n})+K\sum_{e\in W_a(J_l\cup J_r)}|e||\bar{b}_2|^2\\
\leq& Q_a(J)-|a_2b_1|-|a_2b_2|-\sum_{e\in W(J_l\cup J_r)}|ea_2|\\
&+K\sum_{e\in W_a(J_l\cup J_r)}|e|A(\triangle_{m,n})+K\sum_{e\in W_a(J_l\cup J_r)}|e||\bar{b}_2|^2.
\end{aligned}
$$

For Q_d we obtain

$$
\begin{aligned}
Q_d(J')=&\sum_{e_2\in W_a(J_l)}|e_2c_1|+\sum_{e_1\in W_a(J_r)}|e_1c_2|\\
=&\sum_{e_2\in W_a(J_l)}\left|e_2(b_1+O(1)|a_2|+O(1)|\bar{b}_2|^2)\right|\\
&+\sum_{e_1\in W_a(J_r)}\left|e_1(O(1)|a_2|+O(1)|\bar{b}_2|^2)\right|\\
\leq& Q_d(J)-|a_2b_1|-\sum_{e_1\in W_a(J_r)}|e_1a_2|+K\sum_{e\in W_a(J_l\cup J_r)}|e|(|a_2|+|\bar{b}_2|^2).
\end{aligned}
$$

The estimates of Q_{aa} are given by

$$
Q_{aa}(J')=0,\quad Q_{aa}(J)=\frac{1}{2}|a_2|^2+\sum_{e\in W_a(J_l\cup J_r)}|ea_2|.
$$

For Q_{sp} we have

$$
\begin{aligned}
Q_{sp}(\triangle_{J'})&=N_2\{c_1^2+c_2^2\}\\
&\leq Q_{sp}(\triangle_J)+N_2Ka_2^2.
\end{aligned}
$$

Combining the above estimates, we have

$$
\begin{aligned}
Q(J')-Q(J)\leq&-(N_1^2-K-KN_2)|a_2|^2\\
&-(N_1-K-KN_2)\sum_{e\in W(J),e\neq a_2}|e|A(\triangle_{m,n})\\
&-(N_1^2-KN_1)\sum_{e\in W_a(J_l\cup J_r)}|e|A(\triangle_{m,n})\\
&-\{N_2(1-L(J))-3KL(J)\}|\bar{b}_2|^2.
\end{aligned}
$$

From the above inequality, we can easily find N_1 and N_2 satisfying the inequality (3.15). □

From Lemmas 3.5, 3.6, and 3.7, we obtain the following lemma.

LEMMA 3.8. *Suppose J and J' are two consecutive I-curves with $J \le J'$ and $\triangle$ is the diamond between them. Then we have*

$$Q(\triangle) \le 2(Q(J) - Q(J')), \tag{3.16}$$

provided $N_1^2 - K - KN_2 > 0$, $N_1 - K - KN_2 > \frac{1}{2}$, $N_1^2 - KN_1 > 0$, $N_2(1 - L(J)) - 3KL(J) > \frac{1}{2}$, and $L(J) < \frac{1}{2(K+KN_1+KN_1^2)}$.

Let Λ_{J_1,J_2} be the diamonds between J_1 and J_2, and let $A(J_1, J_2)$ be the change in the amount of the approaching waves between J_1 and J_2. The following theorem shows that the total variation of the solution is bounded.

THEOREM 3.9. *Let J_1 and J_2 be the I-curves satisfying $J_1 \le J_2$. Then we have*

$$Q(\Lambda_{J_1,J_2}) \le 2(Q(J_1) - Q(J_2)), \tag{3.17}$$

$$A(J_1, J_2) \le A(J_1) - A(J_2) + KQ(\Lambda_{J_1,J_2}), \tag{3.18}$$

$$L(J_2) - L(J_1) \le K\{A(J_1) - A(J_2) + Q(J_1) - Q(J_2)\}, \tag{3.19}$$

$$P(J_2) - P(J_1) \le K\{A(J_1) - A(J_2) + Q(J_1) - Q(J_2)\}, \tag{3.20}$$

$$L(J) + P(J) \le N_0\{L(J_0) + P(J_0)\}, \tag{3.21}$$

where N_0 is a positive constant. In particular, (3.21) *implies that the global weak solutions exist.*

Proof. The inequality (3.17) is an easy consequence of Lemma 3.8. For (3.18), assume that J and J' are consecutive, $J \le J'$, and $\triangle$ is the diamond between them. If the phase boundary enters $\triangle$, we have

$$A(\triangle) = |a_2| = A(J) - A(J'),$$

and if the phase boundary does not enter $\triangle$, then, assuming as before that P is to the right of $\triangle$, we have

$$\begin{aligned} A(\triangle) &\le |a_2| + |b_2| - |c_2| + KQ(\triangle) \\ &\le A(J) - A(J') + KQ(\triangle). \end{aligned}$$

If the phase boundary enters the diamond between J and J', from Lemma 3.1 we see that, in the case of Lemma 3.6,

$$\begin{aligned} L(J') - L(J) &= |c_1| + |c_2| - |a_2| - |b_1| - |b_2| \\ &\le O(1)|a_2|, \end{aligned}$$

and, in the case of Lemma 3.7,

$$L(J') - L(J) \le O(1)|a_2| + O(1)|\bar{b}_2|^2.$$

Therefore, we have

$$L(J') - L(J) < K\{A(J) - A(J') + Q(J) - Q(J')\}.$$

If the phase boundary does not enter the diamond, from Lemma 3.1 we have that

$$L(J') - L(J) \le K\{Q(J) - Q(J')\}.$$

Repeating this from J_1 to J_2, we obtain (3.19). We can prove (3.20) in a similar manner. Adding (3.19) and (3.20) from J_0 to J, we have

$$\begin{aligned} L(J) + P(J) &\le L(J_0) + P(J_0) + 2K\{A(J_0) + Q(J_0)\} \\ &\le L(J_0) + P(J_0) + 2KA(J_0) + 2K(L(J_0))^2. \end{aligned}$$

From this we obtain (3.21). □

We now define the cancelation. For a diamond into which the phase boundary does not enter,

$$C_i(\triangle) = \frac{1}{2}(|a_i| + |b_i| - |a_i + b_i|).$$

From this we have the following lemma.

LEMMA 3.10. *The following conservation laws hold provided that* Λ *does not contain the phase boundary.*

$$L_i^{\pm}(\Lambda) = E_i^{\pm}(\Lambda) \mp C_i(\Lambda) + O(1)Q(\Lambda), \quad i = 1, 2. \tag{3.22}$$

For a diamond into which the phase boundary enters, assuming that the phase boundary is in the right family of waves entering the diamond, we define

$$C_i(\triangle) = \kappa_i |a_2|,$$

where κ_i is a nonnegative number, so that the following relation holds:

$$L_i(\triangle) = E_i(\triangle) \pm C_i(\triangle), \quad i = 1, 2,$$

where the signs are chosen so that the above equality holds.

The generalized i-characteristics $\chi_i^{\triangle x}(t)$ were introduced in Glimm and Lax [16]. These are the curves consisting of either an i-characteristic or an i-shock issuing from the center of each diamond. They are straight-line segments in each strip $(n-1)\triangle t < t < n\triangle t$ $(n = 0, 1, 2, \ldots)$ and are continued from the center of the diamonds they enter. They showed that $\chi_i^{\triangle x}(t)$ converges uniformly to a Lipschitz function $\chi_i(t)$ as $\triangle x$ approaches zero on any bounded time interval and that the derivatives $\dot{\chi}_i^{\triangle x}(t)$ of the i-generalized characteristics converge pointwise to $\dot{\chi}_i(t)$ except a set of measure zero. Similar results can be obtained for the phase boundary; i.e., the phase boundary $\chi(t)$ is Lipschitz continuous and subsonic, as shown in [4], [26].

4. Initial value problem with the Riemann initial data. We revisit the Riemann problem with the initial data (1.3) and show that the asymptotic states occur immediately for the solution to the Riemann problem.

First, we study the basic estimates concerning the waves in the limit solution. We define the regions Ω_i $(i = 0, 1, 2, 3)$ and Ω_p in the xt-plane as

$$\begin{aligned} \Omega_0 &= \{(x,t) : x < \mu_0 t\}, \\ \Omega_1 &= \{(x,t) : \mu_0 t < x < \mu_1 t\}, \\ \Omega_p &= \{(x,t) : \mu_1 t < x < \mu_2 t\}, \\ \Omega_2 &= \{(x,t) : \mu_2 t < x < \mu_3 t\}, \\ \Omega_3 &= \{(x,t) : \mu_3 t < x\}, \end{aligned}$$

where μ_i $(i = 0, 1, 2, 3)$ satisfy

$$\begin{aligned} \mu_0 &< -\lambda_1(v) - \delta, \\ -\lambda_1(v) + \delta < \mu_1 &< -|\sigma_p| - \delta, \\ |\sigma_p| + \delta < \mu_2 &< \lambda_2(v) - \delta, \\ \lambda_2(v) + \delta &< \mu_3 \end{aligned}$$

for some positive δ and for all of the values of v in the neighborhoods N_{v_α} and N_{v_β}.

Let $\chi_1^1(t)$ and $\chi_1^2(t)$ be the generalized 1-characteristic curves starting from $(\mu_0 t, t)$ and $(\mu_1 t, t)$, respectively. Similarly, define $\chi_2^1(t)$ and $\chi_2^2(t)$ to be the generalized 2-characteristic curves starting from $(\mu_2 t, t)$ and $(\mu_3 t, t)$, respectively. Note that $\chi_i^j(t)$ $(i, j = 1, 2)$ will be redefined in the next section. We define the following:

$$\begin{array}{lcl} \Lambda_i(t) & = & \text{the region between } \chi_i^1(t) \text{ and } \chi_i^2(t), \\ Q(t,s) & = & \text{amount of interaction between } t < t' < s, \\ Q(t) & = & \text{amount of interaction at } t, \\ X_i^{\pm}(s;t) & = & \text{amount of } i\text{-waves in the interior of } \Lambda_i(t) \text{ at time } s, \\ \tilde{X}_i(s;t) & = & \text{amount of } i\text{-waves outside of } \Lambda_i(t) \text{ at time } s, \\ Str.\chi_i^k(s;t) & = & \text{strength of } \chi_i^k(t) \text{ at time } s, \\ D_i(s;t) & = & \text{distance between } \chi_i^1(t) \text{ and } \chi_i^2(t) \text{ at time } s. \end{array} \tag{4.1}$$

Then we obtain the following lemma.

LEMMA 4.1. *The solution to the Cauchy problems* (1.1) *and* (1.3) *converges to the solution to the Riemann problem where the phase boundary does not generate any wave.*

Proof. Suppose that we use an equidistributed sequence $\{\omega_n\}$. In the case of the Riemann initial data, for $t = 0$, we solve a nontrivial Riemann problem in $\triangle_{0,0}$ only, and all of the other Riemann problems are trivial. We solve the same Riemann problem in $\triangle_{0,0}$ for any mesh size. Then, in the next time level, we solve nontrivial Riemann problems in $\triangle_{-1,1}$ and $\triangle_{1,1}$ only, and they do not depend on the mesh size. This implies the following. Suppose that the mesh size is $(\triangle x, \triangle t)$, and we solve a Riemann problem at the mesh point $(m\triangle x, n\triangle t)$. Then, if we change the mesh size to $(\frac{\triangle x}{p}, \frac{\triangle t}{p})$, we solve the above Riemann problem at the mesh point $(\frac{m\triangle x}{p}, \frac{n\triangle t}{p})$ provided that we use the same sequence $\{\omega_n\}$. From the result of the previous section, the Cauchy problem converges. Since $Q(\Lambda_{J_1,J_\infty})$ is finite, all interactions $Q(\triangle_{m,n})$ approach zero as $n \to \infty$. This also implies that the phase boundary is stable in the sense that there are no waves generated from the phase boundary. As $\triangle x$ and $\triangle t$ approach zero maintaining $\frac{\triangle x}{\triangle t} = r$, all of the interactions will take place at the origin, and what remains after taking the limit is the solution where $Q(t)$ is zero for $t > 0$ and all of the waves start from the origin since the phase boundary and all waves of the approximate solutions start from the origin and they are Lipschitz continuous. Since $Q(t) = 0$, the middle states are constant states. Since the waves of the same family interact except when they are both rarefaction waves, either rarefaction waves or shock waves remain in the limiting solutions. Consider the forward wave. We have

$$\frac{dD_i(s;t_o)}{ds} = \sigma_i(U_i^{+2}, U_i^{-2}) - \sigma_i(U_i^{+1}, U_i^{-1}), \tag{4.2}$$

where $i = 2$ and $U_i^{\pm k}$ are the limits of U from the right and left of $\chi_i^k(t)$, $(k = 1, 2)$. Since each characteristic field is genuinely nonlinear, there exists $\theta = \theta(s)$ $(0 < \theta < 1)$

such that

$$\begin{aligned}\frac{dD_i(s;t_o)}{ds} &= \theta(\lambda_i^{-2}(s)-\lambda_i^{+1}(s))+(1-\theta)(\lambda_i^{+2}(s)-\lambda_i^{-1}(s))\\ &= (\lambda_i^{-2}(s)-\lambda_i^{+1}(s))+(1-\theta)(\lambda_i^{+2}(s)-\lambda_i^{-2}(s)+\lambda_i^{+1}(s)-\lambda_i^{-1}(s))\\ &= X_i^+(s;t_o)+X_i^-(s;t_o)+(1-\theta)(Str.\ \chi_i^1(s;t_o)+Str.\ \chi_i^2(s;t_o)).\end{aligned} \tag{4.3}$$

Denote the constant states to the left and right of the phase boundary by U_1^a and U_2^a, respectively. Then, if $\lambda_2(v_R)-\lambda_2(v_2^a)<0$, the forward wave is dominated by shock waves, and, if there are rarefaction waves, they must have interacted with shock waves and canceled out. Therefore, in this case,

$$X_i^+(s;t_o)=0.$$

So, using the above relations and integrating (4.3), we see that

$$D_i(s;t_o)=D_i(t_o;t_o)+\int_{t_o}^{s}\{X_i^-(\tau;t_o)+(1-\theta)(Str.\ \chi_i^1(\tau;t_o)+Str.\ \chi_i^2(\tau;t_o))\}d\tau.$$

Taking the limit as $t_o\to 0$ and noting that $D_i(t_o;t_o)\to 0$, we obtain

$$D_i(s;0)=\int_0^s\{X_i^-(\tau;0)+(1-\theta)(Str.\ \chi_i^1(\tau;0)+Str.\ \chi_i^2(\tau;0))\}d\tau.$$

Note that the integrand is negative, and it is a constant because there is no interaction and cancelation in (3.22). This implies that, if $\lambda_2(v_R)-\lambda_2(v_2^a)<0$ holds, there is a single discontinuity. This discontinuity is a shock because we solve the Riemann problems for the approximate solutions. Since each side of a shock is a constant state, it is a straight line starting from the origin.

If $\lambda_2(v_R)-\lambda_2(v_2^a)>0$, the forward wave is dominated by rarefaction waves, and if there are shock waves, they must have interacted with rarefaction waves and canceled out. Therefore, in this case, for all $\tau>t_o$,

$$X_i^-(\tau;t_o)=0,\ Str.\ \chi_i^1(\tau;t_o)=0,\ Str.\ \chi_i^2(\tau;t_o)=0.$$

Using the conservation laws (3.22) of the rarefaction waves and noting that there is no cancelation, after taking the limit as $t_o\to 0$, we see that

$$D_i(s;0)=sX_i^+(0;0).$$

Since both $Str.\ \chi_i^1(\tau;t_o)$ and $Str.\ \chi_i^2(\tau;t_o)$ are zero, they are the characteristics, and the outsides are constant states. Using the result of Lax [25], we see that the forward wave is a centered rarefaction wave.

If $\lambda_2(v_R)-\lambda_2(v_2^a)=0$, neither is dominant. If both the shocks and rarefactions are present, there should be interactions. If the shocks or rarefaction waves become dominant, we have a contradiction with $\lambda_2(v_R)-\lambda_2(v_2^a)=0$. Therefore, in this case, there is no shock or rarefaction wave for the forward wave. □

The above result implies that the solution to the Riemann problem consists of the constant states U_L, U_1^a, U_2^a, and U_R separated by the backward wave, the phase boundary, and the forward wave, where the middle constant states U_1^a and U_2^a satisfy the following conditions:

$$(4.4)\qquad (1)\qquad \sigma_b \text{ or } -\lambda_1 \le \sigma_p(v_1^a, v_2^a) \le \sigma_f \text{ or } \lambda_2,$$

$$(4.5)\qquad \begin{aligned} u_L + \left\{ \begin{array}{l} -\sigma_b(v_1^a - v_L),\ v_L > v_1^a \\ \int_{v_L}^{v_1^a} \lambda_1(w)dw,\ v_L \le v_1^a \end{array} \right\} \\ -\sigma_p(v_2^a - v_1^a) - \left\{ \begin{array}{l} \sigma_f(v_R - v_2^a),\ v_R < v_2^a \\ \int_{v_2^a}^{v_R} \lambda_2(w)dw,\ v_R \ge v_2^a \end{array} \right\} = u_R; \end{aligned}$$

(2) the admissible solution to the Riemann problem with the initial data U_1^a and U_2^a consists of the phase boundary only.

The following lemma shows that the solution to the Riemann problem satisfying the above conditions exists.

LEMMA 4.2. *If v_L and v_R are close to the Maxwell strains and $u_L \approx u_R$, then U_1^a and U_2^a satisfying* (1) *and* (2) *exist and are unique.*

Proof. Consider case (a) in Figure 2.1. Let the intermediate constant states be $\bar{U}_1 = (\bar{v}_1, \bar{u}_1)$ and $\bar{U}_2 = (\bar{v}_2, \bar{u}_2)$. First, examine condition (2). This condition reduces to the Riemann problem

$$(4.6)\qquad \min\{E_b(v_1^a, \bar{v}_1) + E_p(\bar{v}_1, \bar{v}_2) + E_f(\bar{v}_2, v_2^a)\}$$

subject to

$$(4.7)\qquad \sigma_p A(\bar{v}_1, \bar{v}_2) \le 0,$$

$$(4.8)\qquad \sigma_b \text{ or } -\lambda_1 \le \sigma_p(\bar{v}_1, \bar{v}_2) \le \sigma_f \text{ or } \lambda_2,$$

and

$$(4.9)\qquad \begin{aligned} u_1^a + \left\{ \begin{array}{l} -\sigma_b(\bar{v}_1 - v_1^a),\ v_1^a > \bar{v}_1 \\ \int_{v_1^a}^{\bar{v}_1} \lambda_1(w)dw,\ v_1^a \le \bar{v}_1 \end{array} \right\} \\ -\sigma_p(\bar{v}_2 - \bar{v}_1) - \left\{ \begin{array}{l} \sigma_f(v_2^a - \bar{v}_2),\ v_2^a < \bar{v}_2 \\ \int_{\bar{v}_2}^{v_2^a} \lambda_2(w)dw,\ v_2^a \ge \bar{v}_2 \end{array} \right\} = u_2^a, \end{aligned}$$

where the intermediate states satisfy $\bar{U}_1 = U_1^a$ and $\bar{U}_2 = U_2^a$. Note that $E_b(v_1^a, \bar{v}_1) = O(|v_1^a - \bar{v}_1|^3)$ and $E_f(\bar{v}_2, v_2^a) = O(|\bar{v}_2 - v_2^a|^3)$. First, evaluate $\frac{dE}{d\bar{v}_1}$ at $\bar{v}_1 = v_1^a$. Then

$$(4.10)\qquad \begin{aligned} \left.\frac{dE}{d\bar{v}_1}\right|_{\bar{v}_1 = v_1^a} &= \left.\frac{dE_p}{d\bar{v}_1}\right|_{\bar{v}_1 = v_1^a} \\ &= \left.\frac{1}{4\sigma_p}\left\{\frac{d\bar{v}_2}{d\bar{v}_1}\{\lambda_2^2 - (\sigma_p)^2\}A_{21} - \{\lambda_1^2 - (\sigma_p)^2\}A_{12}\right\}\right|_{\bar{v}_1 = v_1^a}, \end{aligned}$$

where A_{21} and A_{12} are the same as in Theorem 2.5 with the appropriate change of arguments. From (4.9), we see that

$$\left.\frac{d\bar{v}_2}{d\bar{v}_1}\right|_{\bar{v}_1 = v_1^a} = \frac{(\lambda_1^a + \sigma_p^a)^2}{(\lambda_2^a - \sigma_p^a)^2}.$$

Substituting this into (4.10), we have

$$\left.\frac{dE}{d\bar{v}_1}\right|_{\bar{v}_1 = v_1^a} = \frac{\lambda_1^a + \sigma_p^a}{\lambda_2^a - \sigma_p^a}\left\{(\lambda_1^a + \lambda_2^a)A(v_1^a, v_2^a)/(v_2^a - v_1^a) + (v_2^a - v_1^a)(\lambda_1^a\lambda_2^a + (\sigma_p^a)^2)\sigma_p^a\right\},$$

where $\lambda_1^a = \lambda(v_1^a)$, $\lambda_2^a = \lambda(v_2^a)$. If U_1^a is on SC,

$$\left.\frac{dE}{d\bar{v}_1}\right|_{\bar{v}_1 = v_1^a} < 0,$$

and, if U_1^a is on EAC,

$$\left.\frac{dE}{d\bar{v}_1}\right|_{\bar{v}_1 = v_1^a} > 0.$$

Therefore, from the intermediate value theorem and Theorem 2.5, we see that the minimum of E in (4.6) takes place when v_1^a and v_2^a satisfy

$$F_1 = (\lambda_1^a + \lambda_2^a)A(v_1^a, v_2^a)/(v_2^a - v_1^a) + (v_2^a - v_1^a)(\lambda_1^a \lambda_2^a + (\sigma_p^a)^2)\sigma_p^a = 0. \tag{4.11}$$

Also, from (4.9) and the Rankine–Hugoniot condition, we have

$$F_2 = \sigma_p^a(v_2^a - v_1^a) + u_2^a - u_1^a = 0, \tag{4.12}$$

$$F_3 = \sigma_p^a(u_2^a - u_1^a) + f(v_2^a) - f(v_1^a) = 0, \tag{4.13}$$

where U_1^a and U_2^a satisfy

$$F_4 = u_1^a - u_L - \left\{ \begin{array}{l} -\sigma_b(v_1^a - v_L),\ v_1^a \le v_L \\ \int_{v_L}^{v_1^a} \lambda_1(w)dw,\ v_1^a > v_L \end{array} \right\} = 0, \tag{4.14}$$

$$F_5 = u_2^a - u_R - \left\{ \begin{array}{l} \sigma_f(v_R - v_2^a),\ v_R > v_2^a \\ \int_{v_2^a}^{v_R} \lambda_2(w)dw,\ v_R \le v_2^a \end{array} \right\} = 0. \tag{4.15}$$

Hence the problem (4.4)–(4.9) is reduced to (4.11)–(4.15) provided that $v_L \approx v_\alpha$, $v_R \approx v_\beta$, and $u_L \approx u_R$. It can be shown that $U_1^a = (v_\alpha, u_L)$ and $U_2^a = (v_\beta, u_R)$ are the solution if $v_L = v_\alpha$, $v_R = v_\beta$, and $u_L = u_R$, and that the determinant of the Jacobian of $F = (F_1, F_2, F_3, F_4, F_5)$ with respect to $(v_1^a, u_1^a, v_2^a, u_2^a, \sigma_p^a)$ is not zero at $(v_1^a, u_1^a, v_2^a, u_2^a, \sigma_p^a) = (v_L, u_L, v_R, u_L, 0)$. Therefore, the implicit function theorem applies. Case (b) can be shown in a similar manner. □

From the above lemmas, we have the following theorem.

THEOREM 4.3. *The solution to the Riemann problem* (2.3)–(2.6) *with* $U_l = U_L$ *and* $U_r = U_R$ *converges to* (4.11)–(4.15).

Remark 4.1. Since in the Riemann problem the asymptotic states occur immediately after $t = 0$, it is natural to ask if the entropy rate admissibility criterion could have been applied after the separation from the beginning. However, this should be proved, and a proof is provided here. Also, this result does not mean that the criterion can be applied after the separation to all of the Riemann problems. If the strains are specified in the same phase, the hyperbolic solutions and the double phase boundary solutions are two possible solution configurations. In this case, it may be more reasonable to apply the criterion first before the separation to choose the solution configuration and then apply it after the separation to find the stable phase boundaries.

The above result shows that we can solve (4.11)–(4.15) instead of (2.3)–(2.6). In what follows, we apply the entropy rate admissibility criterion after the separation. Then it is not difficult to see that we can go through section 3 to show the existence

of weak solutions. Since the phase boundary is now admissible in each time interval, proceeding in the same manner as [16], we have the following theorem.

THEOREM 4.4. *If η is small, then the weak global solution exists. Furthermore, at the phase boundary, the limit*

$$\lim_{y\to\pm 0} U(\chi(t)+y,t) = U_{\pm}(t) \tag{4.16}$$

exists except for countable t. At the points where the above limit exists, the Riemann problem (4.11)–(4.15) *is solved with the above limits as the initial data. Therefore, the Rankine–Hugoniot conditions are satisfied across the phase boundary, and the entropy condition and the entropy rate admissibility criterion are satisfied at these points.*

5. Large time behavior of solutions. In this section, we study the asymptotic behavior of the solution to (1.1) and (1.2) and compare it with the solution of (1.1) and (1.3). For this purpose, we define ζ_i to be the strength of the i-wave in the solution of the Riemann problem (4.11)–(4.15) with initial data U_L and U_R. We also define $\mathcal{S}$ and $\mathcal{R}$ to be the sets of i's for which the i-wave is a shock and a rarefaction wave, respectively.

Let $\{\chi_L(t),t\}$ and $\{\chi_R(t),t\}$ be the generalized 1- and 2-characteristic curves issuing from $(-M,0)$ and $(M,0)$, respectively, such that $U(x,t)=U_L$ for $x<\chi_L(t)$ and $U(x,t)=U_R$ for $x>\chi_R(t)$. Let $X_i^+(t)$ and $X_i^-(t)$ denote the amount of i-rarefaction and i-shock at time t, respectively. We redefine $\chi_i^1(t)$ and $\chi_i^2(t)$ to be the generalized i-characteristic curves through $\{\chi_L(t),t\}$ and $\{\chi_R(t),t\}$, respectively. Since the phase boundary χ is subsonic, there exists $t_*>t$ such that both $\chi_2^1(t)$ and $\chi_1^2(t)$ finish intersecting with the phase boundary by the time t_* and $t_*=O(1)t$. We set $t_*=t_*(t)$. We also set $T_*=t_*(0)$. For each $t\ge t_*$, $s>t_*$, and $i=1,2$, the above characteristic curves and the phase boundary divide the xt-plane into the following regions:

$$\begin{aligned}
&\Lambda_i(t) = \text{the region between } \chi_i^1(t) \text{ and } \chi_i^2(t),\\
&\Gamma_1(t) = \text{the region between } \chi_2^1(t) \text{ and } \chi_1^2(t) \text{ satisfying } x<\chi(t),\\
&\Gamma_2(t) = \text{the region between } \chi_2^1(t) \text{ and } \chi_1^2(t) \text{ satisfying } x>\chi(t).
\end{aligned}$$

Note that there exists a constant C $(C>1)$ depending only on the system and P_o such that t_* satisfies

$$t_* = Ct. \tag{5.1}$$

We set

$$\begin{aligned}
&X_i(t) = X_i^+(t) + \left|X_i^-(t)\right|, \qquad i=1,2,\\
&X(t) = \sum_{i=1,2} X_i(t).
\end{aligned}$$

In addition to the quantities defined in (4.1), we define

$$\begin{aligned}
A(t,s) &= \text{amount of approaching waves between } t<t'<s,\\
H(t,s) &= \text{amount of } j\text{-waves } (j\neq i) \text{ crossing } \chi_i^1(t) \text{ and } \chi_i^2(t)\\
&\quad \text{between } t<t'<s.
\end{aligned}$$

Then we obtain the following lemma.

LEMMA 5.1. *There exist bounds $O(1)$ such that, for every $s \geq t_*$,*

$$A(s, s') = O(1)Q(t) \text{ for } s' > s, \tag{5.2}$$

$$H(t, s) = O(1)Q(t), \tag{5.3}$$

$$\tilde{X}_i(s; t) = O(1)Q(t), \tag{5.4}$$

$$P(s') - P(s) = O(1)Q(t). \tag{5.5}$$

Proof. Applying the approximate conservation laws to the region outside of $\Lambda_i(t)$, we obtain (5.2), (5.3), and (5.4). From (3.20) we see that

$$|P(s') - P(s)| \leq O(1)A(s, s').$$

This implies (5.5). □

LEMMA 5.2 (Liu [28]). *For $j \in \mathcal{S}$, there exists T_j such that $\chi_i^1(t; t_0)$ and $\chi_i^2(t; t_0)$ impinge to form a shock wave with strength $\zeta_j(t)$. Furthermore, for $t > O(1)|\zeta_j|t > T_j$, we have*

$$|X_j(t) - \zeta_j(t)| \leq Q(O(1)|\zeta_j|t), \qquad j \in \mathcal{S}. \tag{5.6}$$

In what follows, using (3.3) and (3.4), we define

$$Q_s^{\mathcal{R}} = \sum_{i \in \mathcal{R}} O(1)|X_i^-(t)|^3,$$

$$Q_s^{\mathcal{S}} = \sum_{j \in \mathcal{S}} O(1)\zeta_j^2 |X_j(t) - |\zeta_j(t)||,$$

respectively, and then we define

$$Q(t) = Q_s^{\mathcal{R}} + Q_s^{\mathcal{S}} + O(1) \sum_{k=1,2} \eta \tilde{X}_k(t).$$

LEMMA 5.3. *There exist $O(1)$ bounds such that, for every $s \geq t_*$ and for every $(x_1, t_1) \in \Gamma_1(t)$ and $(x_2, t_2) \in \Gamma_2(t)$,*

$$|U(x_1, t_1) - U_1^a| = O(1)Q(O(1)|\zeta_j|t), \tag{5.7}$$

$$|U(x_2, t_2) - U_2^a| = O(1)Q(O(1)|\zeta_j|t). \tag{5.8}$$

Proof. We claim from [10] that there exist constant states $\widetilde{U}_1^a$, $\widetilde{U}_2^a$, and $\widetilde{U}_R$ satisfying

$$\widetilde{U}_1^a \in T_1^r(U_L),\ \widetilde{U}_2^a \in P^r(\widetilde{U}_1^a),\ \widetilde{U}_R \in T_2^r(\widetilde{U}_2^a),$$

and

$$\begin{aligned} &|U(x_1, t_1) - \widetilde{U}_1^a| \\ &= O(1)Q(t) + \sum_{i \in \mathcal{R}} O(1)|X_i^-(t)|^3 + \sum_{j \in \mathcal{S}} O(1)|X_j(t) - |\zeta_j(t)||, \end{aligned} \tag{5.9}$$

$$\begin{aligned} &|U(x_2, t_2) - \widetilde{U}_2^a| \\ &= O(1)Q(t) + \sum_{i \in \mathcal{R}} O(1)|X_i^-(t)|^3 + \sum_{j \in \mathcal{S}} O(1)|X_j(t) - |\zeta_j(t)||, \end{aligned} \tag{5.10}$$

$$|U_R - \widetilde{U}_R| = O(1)Q(t) + \sum_{i \in \mathcal{R}} O(1)|X_i^-(t)|^3 + \sum_{j \in \mathcal{S}} O(1)|X_j(t) - |\zeta_j(t)||, \tag{5.11}$$

where $\widetilde{U}_1^a$ and $\widetilde{U}_2^a$ verify (4.11)–(4.13). From

$$\widetilde{U}_1^a \in T_1^r(U_L)$$

and

$$U(x_1,t_1) \in T_1^r(U_L) + \sum_{i=1,2} O(1)\tilde{X}_i(s;t) + \sum_{i\in\mathcal{R}} O(1)|X_i^-(t)|^3 + \sum_{j\in\mathcal{S}} O(1)|X_j(t) - |\zeta_j(t)||,$$

we obtain the claim (5.9). The other claims are similarly obtained. Since the solution to the Riemann problem (4.4)–(4.9) is differentiable and unique, after moving $\widetilde{U}_R$ to U_R, we obtain the results. □

The following lemma gives the important decay rates of the solutions. Since the proof is lengthy and slightly different from [28], it is postponed until the appendix.

LEMMA 5.4. *Suppose that η is small. Then the Cauchy problem* (1.1), (1.2) *has a global solution satisfying*

$$Q(t) = O(1)t^{-\frac{3}{2}} \text{ as } t \to \infty, \tag{5.12}$$

$$\max_i |X_i^-(t)| = O(1)t^{-\frac{1}{2}}, \quad i \in \mathcal{R}. \tag{5.13}$$

Now we obtain the following result for the asymptotic behavior of solutions.

THEOREM 5.5. *If η is small, then the initial value problem* (1.1), (1.2) *has a global solution satisfying the following estimates as $t \to \infty$. First, in the region outside Λ_i,*

$$\sum_{k=1,2} \tilde{X}_k(t) = O(1)t^{-3/2}. \tag{5.14}$$

If ζ_j is a shock, we have

$$|X_j(t) - \zeta_j(t)| = O(1)t^{-3/2}, \tag{5.15}$$

$$Dist\ \{U(x,t), T_j(U_j^a)\} = O(1)t^{-3/2}, \tag{5.16}$$

where $\zeta_j(t)$ is the strength of the j-shock at t. Furthermore, for $(x_j,t),\ (\bar{x}_j,t) \in \Omega_j$, where (x_j,t) and $(\bar{x}_j,t)$ are on the left and right of the j-shock,

$$|U(x_j,t) - U_{jl}^a| + |U(\bar{x}_j,t) - U_{j_r}^a| = O(1)t^{-3/2}. \tag{5.17}$$

If ζ_i is a rarefaction wave, for $(x,t) \in \Omega_i$, we obtain

$$|U(x,t) - U^a(x,t)| = O(1)t^{-1/2}. \tag{5.18}$$

For $(x_p,t),\ (\bar{x}_p,t) \in \Omega_p$, where (x_p,t) and $(\bar{x}_p,t)$ are on the left and right of the phase boundary, it holds that

$$|U(x_p,t) - U_1^a| + |U(\bar{x}_p,t) - U_2^a| = O(1)t^{-3/2}, \tag{5.19}$$

$$|P(t) - P^a| = O(1)t^{-3/2}, \tag{5.20}$$

where $P(t)$ and P^a are the wave strengths of the phase boundary at t and the phase boundary for the Riemann problem (4.11)–(4.15), *respectively.*

Proof. The proof is similar to Liu [28]. Hence we describe the basic idea of the proof here. The lemmas and theorems referred to from [28] are listed in the appendix. From Theorem 5.4, we see that

$$Q(t) = O(1)t^{-3/2}.$$

Therefore, (5.4) implies that

$$\tilde{X}_i(t) = O(1)Q(t).$$

If $\zeta_j < 0$, it is possible to show that there exists t_0 such that $D_j(t;t_0) = 0$ and $\chi_j^1(t_0)$ and $\chi_j^2(t_0)$ will collide to form shocks with strengths $\zeta_j(t)$. Since outside $\chi_j^1(t_0)$ and $\chi_j^2(t_0)$ the wave strengths of the j-family are $\tilde{X}_j(t)$, (5.15) can be shown from

$$|X_j(t) - |\zeta_j(t)|| = O(1)Q(t) = O(1)t^{-3/2}.$$

From Lemma 5.3 and $\tilde{X}_j(t) = O(1)Q(t)$, we see that (5.17) holds.

For the phase boundary, after $t > T_*$, the strengths of all of the elementary waves entering and leaving are at most $O(1)Q(t)$. From this and Lemma 5.3, we obtain (5.19) and (5.20).

If $\zeta_i > 0$, for (x_i, t) and $(\bar{x}_i, t)$ in Ω_i, (x_j, t) left of $\chi_i^1(t)$, and $(\bar{x}_i, t)$ right of $\chi_i^2(t)$,

$$|u(x_i,t) - u_i^a| + |u(\bar{x}_i,t) - \bar{u}_i^a| = O(1)Q(t^{1/2}) = O(1)t^{-3/4}.$$

Also, from (5.13) we see that the speed of characteristic χ_i in Ω_j for $\tau > t$ is $\lambda_i + O(1)\tau^{-1/2}$. Therefore, the distance $F_i(t)$ between $\chi_i^1(t)$ and the left (or the right) end of the i-rarefaction wave expand at the rate

$$F_i(t) = O(1)t^{1/2}.$$

On the other hand, the centered rarefaction wave is a function of x/t. Therefore, the shift of order $t^{1/2}$ causes the difference of order $t^{-1/2}$. □

Appendix. Here we provide the proof of Lemma 5.4 and list the lemmas used to prove it.

Proof of Lemma 5.4. Since there are several places where the proof is different from [3] and [28], it is presented here. The proof is done by induction. We show that, for some $\bar{T} > 0$, there exist the sequences $\{K_m\}$, $\{L_m\}$, and $\{\gamma_m\}$, $m = 0, 1, 2, \ldots$, such that

$$\max_{i\in\mathcal{R}} |X_i^-(t)| \le K_m t^{-\frac{1}{2}} \quad \text{for} \quad t \le C^m\bar{T}, \tag{A.1}$$

$$Q(t) \le L_m t^{-\frac{3}{2}} \quad \text{for} \quad t \le C^m\bar{T}, \tag{A.2}$$

$$\begin{aligned} K_{m+1} &= K_m[\{1 + O(1)\gamma_m + O(1)K_m^{\rho-1}(C^m\bar{T})^{\xi-\frac{1}{2}} + O(1)K_m^{2\rho-1}(C^m\bar{T})^{2\xi-3/2}\} \\ &\quad \times \left\{1 + O(1)K_m^{\rho}(C^m\bar{T})^{\xi-1}\right\} C^{O(1)\gamma_m} + O(1)K_m^{2-\frac{3\rho}{2}}(C^m\bar{T})^{-\frac{3\xi}{2}+\frac{1}{2}}], \end{aligned} \tag{A.3}$$

$$\gamma_m \le C^{-\delta}\gamma_{m-1} < 1, \tag{A.4}$$

$$L_m = O(1)(K_m)^3, \tag{A.5}$$

where ρ, ξ, and δ are positive numbers chosen later. We show that $K = \overline{\lim}_{m\to\infty} K_m$ and $L = \overline{\lim}_{m\to\infty} L_m$ exist and are finite.

First, we prove (A.1) along with (A.3) and (A.4). For this purpose, we set

$$\bar{T} = (N_0\eta)^{-11/5}\tilde{T}, \quad K_0 = (N_0\eta)^{-1/10}\tilde{T}^{1/2}, \quad L_0 = K_0^3,$$

where $\tilde{T}$ is the time t at which

$$\sum_{j\in\mathcal{S}}\{X_j(t) - |\zeta_j(t)|\} + \sum_{i=1,2}\tilde{X}_i(t) \le \frac{1}{2}\min_{j\in\mathcal{S}}\zeta_j.$$

Note that $\tilde{T}$ may depend on η. Since $K_0\bar{T}^{-\frac{1}{2}} = (N_0\eta)$, (A.1), (A.2), and (A.5) hold for $m = 0$. Now suppose that (A.1)–(A.5) hold for $m = h$. We set

$$T_h = K_h^{\rho}(C^h\bar{T})^{\xi}, \quad \gamma_h = K_h T_h^{-1/2}. \tag{A.6}$$

Note that

$$\gamma_0 = K_0^{(1-\frac{\rho}{2})}(\bar{T})^{-\xi/2} = (N_0\eta)^{-\frac{(1-\frac{\rho}{2})}{10}+\frac{11\xi}{10}}\tilde{T}^{\frac{(1-\frac{\rho}{2})}{2}-\frac{\xi}{2}}.$$

As we will see, we choose $\rho = \frac{6}{5}$, $\xi = \frac{2}{5}$, and $\delta = \frac{3}{20}$ as one choice, and, in this case, we have

$$\gamma_0 = (N_0\eta)^{\frac{19}{50}} \tag{A.7}$$

so that γ_0 is independent of $\tilde{T}$. The induction hypothesis implies that

$$\begin{aligned} &\max_{i\in\mathcal{R}}|X_i^-(t)| \le \gamma_h, \\ &Q(t) \le L_h T_h^{-3/2} = L_h K_h^{-3}\gamma_h^3 = O(1)\gamma_h^3. \end{aligned} \tag{A.8}$$

For $i \in \mathcal{R}$, we have

$$X_i^+(t;T_h) + X_i^-(t;T_h) + Str.\ \chi_i^1(t;T_h) + Str.\ \chi_i^2(t;T_h) = \zeta_i + O(1)\gamma_h^3. \tag{A.9}$$

From this we have

$$|Str.\ \chi_i^1(t;T_h)| + |Str.\ \chi_i^2(t;T_h)| \le X_i^+(t;T_h) - \zeta_i + O(1)\gamma_h^3. \tag{A.10}$$

Using

$$\sigma_i(U_i^{+k}, U_i^{-k}) = \frac{1}{2}(\lambda_i^{+k} + \lambda_i^{-k}) + O(1)|U_i^{+k} - U_i^{-k}|^2, \quad k = 1, 2,$$

in (4.2), we have

$$\begin{aligned} \frac{dD_i(t;T_h)}{dt} &= \frac{1}{2}(\lambda_i^{+2} + \lambda_i^{-2}) + O(1)|U_i^{+2} - U_i^{-2}|^2 \\ &\quad -\frac{1}{2}(\lambda_i^{+1} + \lambda_i^{-1}) + O(1)|U_i^{+1} - U_i^{-1}|^2 \\ &= \lambda_i^{+2} - \lambda_i^{-1} - \frac{1}{2}(\lambda_i^{+2} - \lambda_i^{-2} + \lambda_i^{+1} - \lambda_i^{-1}) \\ &\quad +O(1)|U_i^{+2} - U_i^{-2}|^2 + O(1)|U_i^{+1} - U_i^{-1}|^2 \\ &= \lambda_i^{+2} - \lambda_i^{-1} + \left\{\frac{1}{2} + O(1)|Str.\ \chi_i^1(t;T_h) + Str.\ \chi_i^2(t;T_h)|\right\} \\ &\quad \times \left\{|Str.\ \chi_i^1(t;T_h) + Str.\ \chi_i^2(t;T_h)|\right\}. \end{aligned} \tag{A.11}$$

Since $\lambda_i^{+2} - \lambda_i^{-1} = \lambda_i^{+2}(U_i^a) - \lambda_i^{-1}(U_{i-1}^a) + O(1)\gamma_h^3 = (U_L, U_R)_{r_i} + O(1)\gamma_h^3$, from this and the induction hypotheses (A.1) and (A.8) we have

$$\frac{dD_i(t;T_0)}{dt} = \left(\frac{1}{2} + O(1)\gamma_h\right) K_h t^{-\frac{1}{2}} + \zeta_i + O(1)\gamma_h^3.$$

Integrating this from CT_h to C^hT, we have

$$\begin{aligned} D_i(C^hT;T_h) \le 2\left(\frac{1}{2} + O(1)\gamma_h\right) K_h\{(C^hT)^{1/2} - (CT_h)^{1/2}\} \\ +O(1)\gamma_h^3(C^hT - CT_h) + D_i(CT_h;T_h). \end{aligned} \tag{A.12}$$

Also, from (A.10) and (A.11), we obtain

$$\begin{aligned} \frac{dD_i(t;T_h)}{dt} &\le \left(\frac{1}{2} + O(1)\gamma_h\right)\{X_i^+(t;T_h) - \zeta_i + O(1)\gamma_h^3\} + O(1)\gamma_h^3 \\ &\le \left(\frac{1}{2} + O(1)\gamma_h\right)\left\{\frac{D_i(t;T_h)}{t - CT_h} - \zeta_i\right\} + \zeta_i + O(1)\gamma_h^3. \end{aligned}$$

Integrating this inequality from $C^h\bar{T}$ to t, where $C^h\bar{T} \le t \le C^{h+1}\bar{T}$, and substituting the result into (A.20), we have

$$\begin{aligned} &X_i^+(t) \\ &\le \frac{1}{(t - CT_h)^{\frac{1}{2}}}\left[(1 + O(1)\gamma_h)K_h\frac{\{(C^h\bar{T})^{1/2} - (CT_h)^{1/2}\}}{(C^h\bar{T} - CT_h)^{\frac{1}{2}}} + \frac{D_{r_i}(CT_h;T_h)}{(C^h\bar{T} - CT_h)^{\frac{1}{2}}}\right] \\ &\quad \times\left\{\frac{t - CT_h}{C^h\bar{T} - CT_h}\right\}^{O(1)\gamma_h} + \zeta_i + O(1)\gamma_h^3 \\ &\le t^{-1/2}\left[(1 + O(1)\gamma_h)K_h + \frac{D_{r_i}(CT_h;T_h)}{(C^h\bar{T} - CT_h)^{\frac{1}{2}}}\right] \\ &\quad \times\left(\frac{t}{C^h\bar{T}}\right)^{O(1)\gamma_h}\left\{\frac{1 - \frac{CT_h}{t}}{1 - \frac{CT_h}{C^h\bar{T}}}\right\}^{O(1)\gamma_h}\left(1 - \frac{CT_h}{t}\right)^{-\frac{1}{2}} + \zeta_i + O(1)\gamma_h^3, \end{aligned}$$

where $C^h\bar{T} \le t \le C^{h+1}\bar{T}$. We estimate the terms in the above inequality. For this purpose, we express every term in terms of γ_h, C and show that there exist positive ρ, ξ, and δ satisfying (A.3) and (A.4). From (A.6) we have

$$K_h = \gamma_h^{\frac{1}{1-\frac{\rho}{2}}}(C^h\bar{T})^{\frac{\xi}{2(1-\frac{\rho}{2})}}.$$

The estimates of the terms in the above inequality are given by

$$\frac{CT_h}{C^h\bar{T}} = \frac{CK_h^\rho(C^h\bar{T})^\xi}{C^h\bar{T}} = CK_h^\rho(C^h\bar{T})^{\xi-1} = C\gamma_h^{\frac{\rho}{1-\frac{\rho}{2}}}(C^h\bar{T})^{\frac{\rho\xi}{2(1-\frac{\rho}{2})}}(C^h\bar{T})^{\xi-1},$$

$$\frac{1}{(1-\frac{CT_h}{C^h\bar{T}})^{\frac{1}{2}}} = 1 + O(1)K_h^{\rho}(C^h\bar{T})^{\xi-1},$$

$$\frac{\{(C^h\bar{T})^{1/2} - (CT_h)^{1/2}\}}{(C^h\bar{T} - CT_h)^{\frac{1}{2}}} \le 1,$$

$$\frac{D_i(CT_h;T_h)}{(C^h\bar{T}-CT_h)^{\frac{1}{2}}} = \frac{O(1)T_h}{(C^h\bar{T}-CT_h)^{\frac{1}{2}}} = \frac{O(1)K_h^{\rho}(C^h\bar{T})^{\xi-\frac{1}{2}}}{(1-\frac{CT_h}{C^h\bar{T}})^{\frac{1}{2}}}$$
$$\le O(1)K_h^{\rho}(C^h\bar{T})^{\xi-\frac{1}{2}}\{1 + O(1)K_h^{\rho}(C^h\bar{T})^{\xi-1}\}.$$

From (A.7), $O(1)\gamma_0 < \frac{1}{2}$ is possible if we choose η to be small. Assuming that $O(1)\gamma_h < \frac{1}{2}$ is possible as a part of the induction hypothesis, we see that

$$\begin{aligned}
&\left\{\frac{1-\frac{CT_h}{t}}{1-\frac{CT_h}{C^h\bar{T}}}\right\}^{O(1)\gamma_h}\left(1-\frac{CT_h}{t}\right)^{-\frac{1}{2}}\\
&\le \left(1-\frac{CT_h}{C^h\bar{T}}\right)^{-(\frac{1}{2}+O(1)\gamma_h)}\\
&= \{1-O(1)(\gamma_h)^{\frac{\rho}{1-\frac{\rho}{2}}}(C^h\bar{T})^{\frac{\rho\xi}{2(1-\frac{\rho}{2})}}(C^h\bar{T})^{\xi-1}\}^{-(\frac{1}{2}+O(1)\gamma_h)}\\
&\le 1 + O(1)(\gamma_h)^{\frac{\rho}{1-\frac{\rho}{2}}}(C^h\bar{T})^{\frac{\rho\xi}{2(1-\frac{\rho}{2})}}(C^h\bar{T})^{\xi-1}.
\end{aligned}$$

Hence

$$\begin{aligned}
&X_i^+(t)\\
&\le K_h t^{-1/2}[\{1 + O(1)\gamma_h + O(1)K_h^{\rho-1}(C^h\bar{T})^{\xi-\frac{1}{2}}(1+O(1)K_h^{\rho}(C^h\bar{T})^{\xi-1})\}\\
&\quad\times\left\{1+O(1)K_h^{\rho}(C^h\bar{T})^{\xi-1}\right\}C^{O(1)\gamma_h} + O(1)K_h^{2-\frac{3\rho}{2}}(C^h\bar{T})^{-\frac{3\xi}{2}+\frac{1}{2}}] + \zeta_i.
\end{aligned}$$

We define

$$\begin{aligned}
&K_{h+1}\\
&= K_h[\{1 + O(1)\gamma_h + O(1)K_h^{\rho-1}(C^h\bar{T})^{\xi-\frac{1}{2}}(1+O(1)K_h^{\rho}(C^h\bar{T})^{\xi-1})\}\\
&\quad\times\left\{1+O(1)K_h^{\rho}(C^h\bar{T})^{\xi-1}\right\}C^{O(1)\gamma_h} + O(1)K_h^{2-\frac{3\rho}{2}}(C^h\bar{T})^{-\frac{3\xi}{2}+\frac{1}{2}}]
\end{aligned} \tag{A.13}$$

and examine the inequality (A.4).

$$\begin{aligned}
\frac{\gamma_{h+1}}{\gamma_h} &= \left(\frac{K_{h+1}}{K_h}\right)^{(1-\frac{\rho}{2})}C^{-\xi/2}\\
&= [\{1 + O(1)\gamma_h + O(1)\gamma_h^{\frac{\rho-1}{1-\frac{\rho}{2}}}(C^h\bar{T})^{\frac{\xi(\rho-1)}{2(1-\frac{\rho}{2})}}(C^h\bar{T})^{\xi-\frac{1}{2}}\\
&\quad\times(1+O(1)C\gamma_h^{\frac{\rho}{1-\frac{\rho}{2}}}(C^h\bar{T})^{\frac{\rho\xi}{2(1-\frac{\rho}{2})}}(C^h\bar{T})^{\xi-1})\}\\
&\quad\times\{1+O(1)C\gamma_h^{\frac{\rho}{1-\frac{\rho}{2}}}(C^h\bar{T})^{\frac{\rho\xi}{2(1-\frac{\rho}{2})}}(C^h\bar{T})^{\xi-1}\}C^{O(1)\gamma_h}\\
&\quad+\gamma_h^{\frac{2-\frac{3\rho}{2}}{1-\frac{\rho}{2}}}(C^h\bar{T})^{\frac{\xi(2-\frac{3\rho}{2})}{2(1-\frac{\rho}{2})}}(C^h\bar{T})^{-\frac{3\xi}{2}+\frac{1}{2}}]^{(1-\frac{\rho}{2})}C^{-\xi/2}.
\end{aligned} \tag{A.14}$$

Assuming that the induction hypothesis $\gamma_h < C^{-\delta h}\gamma_0$ holds, we obtain the condition under which $\frac{\gamma_{h+1}}{\gamma_h} < C^{-\delta}$ is satisfied. After substituting $\gamma_h < C^{-\delta h}\gamma_0$ in (A.14), we see that, if the following inequality is satisfied, $\frac{\gamma_{h+1}}{\gamma_h} < C^{-\delta}$ is satisfied.

$$
\begin{aligned}
(A.15)\quad & [\{1+O(1)C^{-\delta h}\gamma_0 \\
& +O(1)C^{\{\frac{\rho-1}{1-\frac{\rho}{2}}(\frac{\xi}{2}-\delta)+\xi-\frac{1}{2}\}h}(\gamma_0)^{\frac{\rho-1}{1-\frac{\rho}{2}}}(\bar{T})^{\frac{\xi(\rho-1)}{2(1-\frac{\rho}{2})}}(\bar{T})^{\xi-\frac{1}{2}} \\
& \times(1+O(1)C^{\{\frac{\rho}{1-\frac{\rho}{2}}(\frac{\xi}{2}-\delta)+\xi-1\}h}(\gamma_0)^{\frac{\rho}{1-\frac{\rho}{2}}}(\bar{T})^{\frac{\rho\xi}{2(1-\frac{\rho}{2})}}(\bar{T})^{\xi-1})\} \\
& \times\{1+O(1)C^{\{\frac{\rho}{1-\frac{\rho}{2}}(\frac{\xi}{2}-\delta)+\xi-1\}h}(\gamma_0)^{\frac{\rho}{1-\frac{\rho}{2}}}(\bar{T})^{\frac{\rho\xi}{2(1-\frac{\rho}{2})}}(\bar{T})^{\xi-1}\}C^{O(1)C^{-\delta h}\gamma_0} \\
& +O(1)C^{\{\frac{2-\frac{3\rho}{2}}{1-\frac{\rho}{2}}(\frac{\xi}{2}-\delta)-\frac{3\xi}{2}+\frac{1}{2}\}h}(\gamma_0)^{\frac{2-\frac{3\rho}{2}}{1-\frac{\rho}{2}}}(\bar{T})^{\frac{\xi(2-\frac{3\rho}{2})}{2(1-\frac{\rho}{2})}}(\bar{T})^{-\frac{3\xi}{2}+\frac{1}{2}]^{(1-\frac{\rho}{2})}} \\
& < C^{-\delta+\frac{\xi}{2}}.
\end{aligned}
$$

Examining the terms in (A.15), we see that a sufficient condition for (A.15) is that the power of C^h is negative, the power of $(N_0\eta)$ is positive, and the power of $\tilde{T}$ is nonpositive. To make the power of C^h negative, we take $\frac{1}{3} < \xi < \frac{1}{2}$, $\frac{1}{3} < \rho < \frac{4}{3}$, and $\frac{\xi}{2} - \delta > 0$. Observe that the powers of $\tilde{T}$ in the above expressions are proportional to $(\rho + 2\xi - 2)$. We choose ρ and ξ so that $\rho + 2\xi - 2 = 0$. One example satisfying the above restrictions is $\rho = \frac{6}{5}$, $\xi = \frac{2}{5}$, and $\delta = \frac{3}{20}$. Then we have

$$
X_i^+(t) \le K_{h+1}t^{-\frac{1}{2}} + \zeta_i, \quad i \in \mathcal{R}, \tag{A.16}
$$

and we obtain from (A.13) the following estimate for the sequence $\{K_m\}$:

$$
K_h < (1+O(1)C^{-\nu h})^h C^{O(1)\sum_h \gamma_h} K_0,
$$

where ν is a positive constant. This shows that $K = \overline{\lim}_{m\to\infty} K_m$ exists and is finite. Therefore, we see that the estimates (A.1), (A.3), and (A.4) hold for $m = h+1$.

Next, we prove (A.2) and (A.5). From (A.21), (A.1), and (A.4), we have for $m \le h$

$$
\begin{aligned}
Q(t) &= Q_s(t) + \sum_{k=1}^{h}(O(1)\eta)^k Q_s(C^{-k}t) + (O(1)\eta)^h Q_d(C^{-p}t) \\
&= Q_s(t) + O(1)\sum_{k=1}^{h}(O(1)\eta)^k K_h^3 (C^{-k}t)^{-3/2}, \quad C^h\bar{T} \le t \le C^{h+1}\bar{T}. \qquad (A.17)
\end{aligned}
$$

The estimate of $Q_s(t)$ is made in the following way. For $i \in \mathcal{R}$, from (A.1) and (A.16), we have for $m \le h+1$

$$
Q_s^{\mathcal{R}}(t) = \sum_{i\in\mathcal{R}} |X_i^-(t)|^3 = O(1)K_{h+1}^3 t^{-3/2}. \tag{A.18}
$$

For $j \in \mathcal{S}$, there exists $t' < t$ such that $x_j^1(t')$ and $x_j^2(t')$ meet before time t. Since $t \le O(1)t'/|\zeta_j|$ and $K_h < K_{h+1}$,

$$
\begin{aligned}
Q_s^{\mathcal{S}}(t) &= \sum_{j\in\mathcal{S}} O(1)\zeta_j^2 |X_j(t) - \zeta_j(t)| \\
&= \sum_{j\in\mathcal{S}} O(1)\zeta_j^2 Q(t') \\
&= \sum_{j\in\mathcal{S}} O(1)|\zeta|_j^{1/2} K_{h+1}^3 t^{-3/2}. \qquad (A.19)
\end{aligned}
$$

Therefore, (A.17), (A.18), and (A.19) imply (A.2) and (A.5). □

LEMMA A.1 (Glimm and Lax [16]). *For a genuinely nonlinear i-characteristic family and $\tau \geq t > t_*$, we have*

$$X_i^+(\tau;t) \leq \frac{D_i(\tau;t)}{\tau - t_*} + O(1)Q(t). \tag{A.20}$$

LEMMA A.2 (Liu [28]). *For the constant C defined in* (5.1),

$$Q(C^m t) = \sum_{k=0}^{m} (O(1)\eta)^k Q_s(C^{m-k}t) + (O(1)\eta)^m Q_d(t), \tag{A.21}$$

where η is defined as in (1.4).

REFERENCES

[1] R. ABEYARATNE AND J. K. KNOWLES, *Kinetic relations and the propagation of phase boundaries in solids*, Arch. Ration. Mech. Anal., 114 (1991), pp. 119–154.

[2] R. ABEYARATNE AND J. K. KNOWLES, *On the propagation of maximally dissipative phase boundaries in solids*, Quart. Appl. Math., 50 (1992), pp. 149–172.

[3] F. ASAKURA, *Asymptotic stability of solutions with a single strong shock wave for hyperbolic conservation laws*, Japan J. Indust. Appl. Math., 11 (1994), pp. 225–244.

[4] F. ASAKURA, *Large time stability of the Maxwell states*, Methods Appl. Anal., 6 (1999), pp. 477–503.

[5] I. L. CHERN, *Stability theorem and truncation error analysis for the Glimm scheme and for a front tracking method for flows with strong discontinuities*, Comm. Pure Appl. Math., 42 (1989), pp. 815–844.

[6] R. M. COLOMBO AND A. CORLI, *Continuous dependence in conservation laws with phase transitions*, SIAM J. Math. Anal., 31 (1999), pp. 34–62.

[7] A. CORLI AND M. SABLÉ-TOUGERON, *Kinetic stabilization of nonlinear sonic phase boundary*, Arch. Ration. Mech. Anal., 152 (2000), pp. 1–63.

[8] C. M. DAFERMOS, *The entropy rate admissibility criterion for solutions of hyperbolic conservation laws*, J. Differential Equations, 14 (1973), pp. 202–212.

[9] C. M. DAFERMOS, *The entropy rate admissibility criterion in thermoelasticity*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 8 (1974), pp. 113–119.

[10] R. J. DIPERNA, *Decay of solutions of hyperbolic systems of conservation laws with a convex extension*, Arch. Ration. Mech. Anal., 64 (1977), pp. 1–46.

[11] H. FAN, *A vanishing viscosity approach on the dynamics of phase transitions in van der Waals fluid*, J. Differential Equations, 103 (1993), pp. 179–204.

[12] H. FAN, *A limiting "viscosity" approach to the Riemann problem for the materials exhibiting change of phase*, Arch. Ration. Mech. Anal., 116 (1992), pp. 317–337.

[13] H. FAN, *The uniqueness and stability of the solution of the Riemann problem of a system of conservation laws of mixed type*, Trans. Amer. Math. Soc., 333 (1992), pp. 913–938.

[14] H. FAN AND M. SLEMROD, *The Riemann problem for systems of conservation laws of mixed type*, in Shock Induced Transitions and Phase Structures in General Media, IMA Vol. Math. Appl. 52, Springer-Verlag, New York, 1993, pp. 61–91.

[15] J. GLIMM, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[16] J. GLIMM AND P. D. LAX, *Decay of solutions of systems of nonlinear hyperbolic conservation laws*, Mem. Amer. Math. Soc., 101 (1970).

[17] H. HATTORI, *The Riemann problem for a van der Waals fluid with entropy rate admissibility criterion: Isothermal case*, Arch. Ration. Mech. Anal., 92 (1986), pp. 246–263.

[18] H. HATTORI, *The Riemann problem for a van der Waals fluid with entropy rate admissibility criterion: Non-isothermal case*, J. Differential Equations, 65 (1986), pp. 158–174.

[19] H. HATTORI, *The entropy rate admissibility criterion and the entropy condition for a phase transition problem: The isothermal case*, SIAM J. Math. Anal., 31 (2000), pp. 791–820.

[20] H. HATTORI AND K. MISCHAIKOW, *A dynamical systems approach to a phase transition problem*, J. Differential Equations., 94 (1991), pp. 340–378.

[21] D. Hoff and M. Khodja, *Stability of coexisting phases for compressible van der Waals fluids*, SIAM J. Appl. Math., 53 (1993), pp. 1–14.
[22] L. Hsiao, *Uniqueness of admissible solutions of the Riemann problem for a system of conservation laws of mixed type*, J. Differential Equations., 86 (1990), pp. 197–233.
[23] J. D. James, *The propagation of phase boundaries in elastic bars*, Arch. Ration. Mech. Anal., 73 (1980), pp. 125–158.
[24] B. L. Keyfitz, *The Riemann problem for nonmonotone stress-strain functions: A "hysteresis" approach*, in Nonlinear Systems of Partial Differential Equations in Applied Mathematics, Part 1, Lectures in Appl. Math. 23, AMS, Providence, RI, 1986, pp. 379–395.
[25] P. D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1967), pp. 537–556.
[26] P. Le Floch, *Propagating phase boundaries: Formulation of the problem and existence via Glimm's scheme*, Arch. Ration. Mech. Anal., 123 (1993), pp. 153–197.
[27] T. P. Liu, *Decay to N-waves of solutions of general systems of nonlinear hyperbolic conservation laws*, Comm. Pure Appl. Math., 30 (1977), pp. 585–610.
[28] T. P. Liu, *Linear and nonlinear large time behavior of solutions of general systems of hyperbolic conservation laws*, Comm. Pure Appl. Math., 30 (1977), pp. 767–796.
[29] J. M. Mercier and B. Piccoli, *Global continuous Riemann solver for nonlinear elasticity*, Arch. Ration. Mech. Anal., 156 (2001), pp. 89–119.
[30] P. L. Pego, *Phase transitions in one-dimensional nonlinear viscoelasticity: Admissibility and stability*, Arch. Ration. Mech. Anal., 97 (1987), pp. 353–394.
[31] R. L. Pego and D. Serre, *Instabilities in Glimm's scheme for two systems of mixed type*, SIAM J. Numer. Anal., 25 (1988), pp. 965–988.
[32] T. J. Pence, *On the mechanical dissipation of solutions to the Riemann problem for impact involving a two-phase elastic material*, Arch. Ration. Mech. Anal., 117 (1992), pp. 1–55.
[33] M. Shearer, *The Riemann problem for a class of conservation laws of mixed type*, J. Differential Equations., 46 (1982), pp. 426–443.
[34] M. Shearer, *Nonuniqueness of admissible solutions of the Riemann initial value problem for a system of conservation laws of mixed type*, Arch. Ration. Mech. Anal., 93 (1986), pp. 45–59.
[35] M. Slemrod, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Ration. Mech. Anal., 81 (1983), pp. 301–315.
[36] M. Slemrod, *Dynamic phase transitions in a van del Waals fluid*, J. Differential Equations, 52 (1984), pp. 1–23.
[37] M. Slemrod, *A limiting "viscosity" approach to the Riemann problem for materials exhibiting change of phase*, Arch. Ration. Mech. Anal., 105 (1989), pp. 327–365.
[38] L. I. Slepyan, *Principle of maximum energy dissipation rate in crack dynamics*, J. Mech. Phys. Solids, 41 (1993), pp. 1019–1033.

SIAM J. MATH. ANAL.
Vol. 34, No. 4, pp. 805–835

ON COMPLEX-VALUED SOLUTIONS TO A TWO-DIMENSIONAL EIKONAL EQUATION. II. EXISTENCE THEOREMS*

ROLANDO MAGNANINI† AND GIORGIO TALENTI†

Abstract. The equation $w_x^2 + w_y^2 + n^2(x,y) = 0$, which arises in generalizations of geometrical optics, is investigated from a theoretical point of view. Here x and y denote rectangular coordinates in the Euclidean plane, and n is real-valued and strictly positive. A framework is set up that involves a Bäcklund transformation relating $\mathrm{Re}(w)$ and $\mathrm{Im}(w)$, second-order partial differential equations in divergence and nondivergence form governing $\mathrm{Re}(w)$, a variational integral, and related free boundary problems, boundary value problems, and viscosity solutions. The present paper is a continuation of a preceding one [R. Magnanini and G. Talenti, *Contemp. Math.* 283, AMS, Providence, RI, 1999, pp. 203–229], where qualitative properties of smooth solutions are offered. Here the existence of the real part of solutions, which need not be smooth, is derived.

Key words. partial differential equations, Bäcklund transformations, convex functionals, minimizers, free boundaries, critical points, variational solutions, viscosity solutions

AMS subject classifications. Primary, 35J70, 35Q60; Secondary, 49N60

PII. S0036141002400877

1. Introduction.

1.1. General. Let x and y denote rectangular coordinates in the Euclidean plane $\mathbb{R}^2$, and let n be a *real-valued* function of x and y. Let n be sufficiently smooth and strictly positive; should the range of x and y be unbounded, let n decay fast enough at infinity. The present paper, its predecessor [26], and forthcoming others are devoted to a tentative theory of the partial differential equation

$$\left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2 + n^2(x,y) = 0, \tag{1.1}$$

all of whose solutions are *complex-valued.*

Versions of (1.1) arise in acoustics and optics. Suppose that a two-dimensional isotropic nondissipative medium is under consideration and that n represents the relevant refractive index. Since (1.1) turns into

$$w_x^2 + w_y^2 = n^2(x,y)$$

on replacing w by $\pm iw$, the solutions to (1.1) whose real part is zero call for processes of classical geometrical optics. (We denote $\sqrt{-1}$ by i throughout and denote differentiations either by $\partial/\partial x$ and $\partial/\partial y$ or by subscripts.) On the other hand, solutions to (1.1) whose real part is different from zero are alleged to account for an optical process that is inherently excluded from geometrical optics—the development of *evanescent waves.* Evanescent waves occur beyond a caustic, on the dark side where the geometric optical rays do not penetrate, or else on the optically thinner side of

*Received by the editors April 25, 2001; accepted for publication (in revised form) July 19, 2002; published electronically February 6, 2003. This work was supported by a 1999–2000 grant of the Italian MURST.

http://www.siam.org/journals/sima/34-4/40087.html

†Dipartimento di Matematica U. Dini, Università di Firenze, viale Morgagni 67/A, 50134 Firenze, Italy (magnanin@math.unifi.it, talenti@math.unifi.it).

an interface that disconnects two different media and totally reflects a wave incident from the optically denser side. A theory, put forward by Felsen and coworkers some twenty years ago and sometimes called evanescent wave tracking (EWT), claims that features of evanescent waves can be portrayed by retaining the asymptotic expansion

$$\text{electromagnetic field} \sim \exp[-i\nu \cdot (\text{time})] \cdot (\text{amplitude}) \cdot \exp[i\nu \cdot (\text{eikonal})],$$

which lies at the very root of geometrical optics, but allowing the eikonal and the components of the amplitude to take *complex values*; here the amplitude and the eikonal are functions of space coordinates only, and ν, the wave number, tends to infinity. A key to EWT amounts precisely to (1.1) and its three-dimensional analogue. By the way, these same objects appear also in a more exhaustive asymptotic analysis of the electromagnetic field, which leads to uniform expansions near caustics, and in modeling deeper diffraction processes. More information can be found in [4], [5], [10], [11], [14], [15], [18], [20], [21], [24], [25], [23], and in the recent surveys [3] and [6].

1.2. Preparatory results. We warm up by recollecting some material from [26]. Let u and v be *real-valued* functions of x and y, and let

$$w = u + iv$$

be the *complex-valued* function of x and y whose real and imaginary parts are u and v, respectively. w is a solution to (1.1) if and only if u and v obey the following system:

$$\begin{aligned} u_x^2 + u_y^2 - v_x^2 - v_y^2 + n^2 &= 0, \\ u_x v_x + u_y v_y &= 0. \end{aligned} \tag{1.2}$$

u and v obey (1.2) if and only if either

$$u_x = u_y = 0 \quad \text{and} \quad v_x^2 + v_y^2 = n^2$$

or the condition

$$u_x^2 + u_y^2 > 0$$

and the following equations

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = \pm\sqrt{1 + \frac{n^2}{u_x^2 + u_y^2}} \begin{bmatrix} -u_y \\ u_x \end{bmatrix}, \tag{1.3}$$

$$\frac{\partial}{\partial x}\left\{\sqrt{1 + \frac{n^2}{u_x^2 + u_y^2}}\, u_x\right\} + \frac{\partial}{\partial y}\left\{\sqrt{1 + \frac{n^2}{u_x^2 + u_y^2}}\, u_y\right\} = 0 \tag{1.4}$$

prevail.

Equations (1.3), which result from algebraic manipulations of (1.2), define a *Bäcklund transformation.* (An account of Bäcklund transformations which fits well into the present context is in [30].) Equation (1.4), which amounts to the integrability of (1.3), is a second-order partial differential equation in *divergence form.* If sufficiently smooth solutions are considered whose gradient is different from 0, (1.4) can be recast in the form

$$\begin{aligned} &[(u_x^2 + u_y^2)^2 + n^2 u_y^2] u_{xx} - 2n^2 u_x u_y u_{xy} + [(u_x^2 + u_y^2)^2 + n^2 u_x^2] u_{yy} \\ &\qquad + n\left(u_x^2 + u_y^2\right)(n_x u_x + n_y u_y) = 0, \end{aligned} \tag{1.5}$$

a *semilinear* second-order partial differential equation with *polynomial nonlinearities.* Equations (1.4) and (1.5) are *elliptic-parabolic* or *degenerate elliptic.* A real-valued solution u to either (1.4) or (1.5) is *elliptic* if $u_x^2 + u_y^2 > 0$; a *degeneracy* occurs at any point where $u_x = u_y = 0$.

It should be stressed that (1.4) and (1.5) are *not* equivalent. First, perfectly smooth solutions to (1.5) exist, whose gradients vanish exclusively in a set of measure 0, and that *do not* satisfy (1.4) in the sense of distributions; they make the left-hand side (l.h.s.) of (1.4) a well-defined distribution which is supported by the set of the critical points but is *not* zero. The identity

$$\text{l.h.s. of (1.5)} = (n^2 + u_x^2 + u_y^2)^{\frac{1}{2}}(u_x^2 + u_y^2)^{\frac{3}{2}} \times \{ \text{ l.h.s. of (1.4)}\}$$

gives evidence to such a statement. In the case in which $n \equiv 1$, one of the last mentioned solutions is constructed by selecting a constant C such that $0 < C < 1$ (e.g., $C = 10^{-10}$) and letting

$$\begin{aligned} &\text{domain of } u = \{(x,y) : x^2/(1-C^2) - y^2/C^2 < 1\}, \\ &\sqrt{2}\cdot u(x,y) = (((1-x^2-y^2)^2 + 4y^2)^{1/2} + 1 - x^2 - y^2)^{1/2}; \end{aligned} \tag{1.6}$$

see [26, Proposition 2.2.1]. Second, we shall demonstrate in the present paper that a conventional boundary condition need not determine a solution to (1.4) in the whole of a domain prescribed in advance, whereas the same boundary condition does suit appropriate solutions to (1.5).

The two theorems below, which bring *critical points* into relation with *rays,* express distinctive properties of the equations in hand. Recall the following. A point where the gradient vanishes is qualified as *critical.* A critical point where the Hessian determinant vanishes is qualified as *degenerate.* (The implicit function theorem states that the gradient of a sufficiently smooth real-valued function acts as a diffeomorphism from a neighborhood of a nondegenerate critical point into a neighborhood of the origin. Therefore, any nondegenerate critical point is isolated, and, conversely, all nonisolated critical points are degenerate.) The geodesics belonging to the Riemannian metric

$$n(x,y)\sqrt{(dx)^2 + (dy)^2}, \tag{1.7}$$

i.e., the paths making

$$\int n(x,y)\sqrt{(dx/ds)^2 + (dy/ds)^2}\, ds$$

either stationary or a minimum, are nicknamed *rays* and are characterized by the differential equation

$$(\text{gradient of } \log n)\cdot(\text{principal normal}) = 1. \tag{1.8}$$

THEOREM 1.1. *Assume n is strictly positive and w is a smooth solution to* (1.1). *If the gradient of* $\mathrm{Re}(w)$ *vanishes at some point, then the same gradient vanishes everywhere on a ray passing through that point.*

THEOREM 1.2. *Suppose n is smooth and strictly positive. Suppose u is smooth and real-valued and satisfies either* (1.4) *or* (1.5) *in every open subset of its domain where $u_x^2 + u_y^2 > 0$. We make the following assertions:*

(i) *Any critical point of u is degenerate.*

(ii) *If* $u_x = u_y = 0$ *and* $u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2 > 0$ *at some point, then* $u_x = u_y = 0$ *everywhere on a smooth curve passing through that point.*

(iii) *If* $u_x = u_y = 0$ *and* $u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2 > 0$ *at every point of a smooth curve, then this curve is a ray.*

Theorem 1.1 makes arguments from [14] rigorous. It also offers a proof of the following statement, which plays a role in the so-called *theory of complex rays* and was alleged in [11, section 3.2]. Let w be a solution to (1.1); if a point obeys the principle of locality, i.e., is a critical point of $\mathrm{Re}(w)$, then the phase path crossing that point, i.e., the level line of $\mathrm{Re}(w)$ containing the point in question, is a ray.

Theorem 1.2 basically shows that (1.5), unlike more conventional second-order partial differential equations, prevents its solutions from having *isolated critical points.* The degeneracy at critical points is a feature of (1.5) that causes critical points to cluster.

Another relevant feature is the architecture of (1.5), which exhibits geometric ingredients. If critical points are ignored and h is defined by either

$$h = -\left(u_x^2 + u_y^2\right)^{-3/2}\left(u_y^2 u_{xx} - 2u_x u_y u_{xy} + u_x^2 u_{yy}\right)$$

or

$$h = -\mathrm{div}\left(\frac{\nabla u}{|\nabla u|}\right),$$

then (1.5) reads both

$$|\nabla u|\ \Delta u - n^2\left\{h - \nabla \log n \cdot \frac{\nabla u}{|\nabla u|}\right\} = 0$$

and

$$\left(\frac{u_x}{|\nabla u|}\frac{\partial}{\partial x} + \frac{u_y}{|\nabla u|}\frac{\partial}{\partial y}\right)\log\sqrt{n^2 + |\nabla u|^2} = h.$$

(We denote the divergence operator by div and the gradient operator by ∇. We denote the length of a vector by vertical bars and the scalar product of two vectors by either a dot or parentheses. For instance, we let

$$|\nabla u| = \sqrt{u_x^2 + u_y^2} \quad \text{and} \quad \nabla u \cdot \nabla v = (\nabla u, \nabla v) = u_x v_x + u_y v_y$$

in case that u and v are real-valued. As usual,

$$\Delta = \partial^2/\partial x^2 + \partial^2/\partial^2 y,$$

the Laplace operator.) Observe the following. First, the principal normal to the level lines of u is

$$(1/h)\frac{\nabla u}{|\nabla u|};$$

in other words, the value of h at any point (x, y) is a *signed curvature* at (x, y) of the *level line* of u crossing (x, y). Second, the value of

$$\nabla \log n \cdot \frac{\nabla u}{|\nabla u|}$$

at (x, y) equals a *signed curvature* at (x, y) of the *ray* which is tangent at (x, y) to a level line of u. Third,

$$\frac{u_x}{|\nabla u|}\frac{\partial}{\partial x} + \frac{u_y}{|\nabla u|}\frac{\partial}{\partial y}$$

is a directional derivative along the *lines of steepest descent* of u. (The first statement follows from Frenet's formulas; the second statement is a consequence of the differential equation (1.8), which characterizes rays; the third one amounts to saying that the lines of steepest descent are the trajectories of the gradient.)

1.3. Background. The present paper rests upon a background that we fix now. We borrow terminology from [1], [28], [29], [31], [33], [34], and the theory of distributions and offer apropos details in the next paragraphs.

Equation (1.4) is reminiscent of the *Euler–Lagrange equation* of a variational integral. Let

$$\Omega = \text{some open nonempty subset of } \mathbb{R}^2, \tag{1.9}$$

let a real function f be defined by

$$f(\rho) = \frac{1}{2}[\rho\sqrt{\rho^2+1} + \log(\rho + \sqrt{\rho^2+1})] \tag{1.10}$$

for every nonnegative ρ, and let a functional J be defined by

$$J(u) = \int_\Omega f\left(\frac{|\nabla u|}{n}\right) n^2 dxdy \tag{1.11}$$

for every u from some set of nice real-valued functions of x and y.

Observe that, if the Riemannian metric (1.7) is in force, the expressions

$$\frac{|\nabla u|}{n} \quad \text{and} \quad n^2 dxdy,$$

which appear in (1.11), equal the Riemannian length of the covariant derivative of u and the Riemannian area element, respectively. As will be clear presently, the right-hand side (r.h.s.) of (1.11) would become the Riemannian area of the graph of u if f were replaced by its derivative f'.

Equation (1.10) gives $f(0) = 0$,

$$f'(\rho) = \sqrt{\rho^2+1},$$

and

$$f''(\rho) = \rho/f'(\rho)$$

for every nonnegative ρ; moreover, $f''' = (f')^{-3}$. We infer that f is nonnegative, vanishes only at 0, and is strictly increasing and strictly convex—a good *Young function.* Therefore, functional J is *strictly convex,* provided a convex domain is supplied to it.

Roughly, a domain that fits J well consists of real-valued functions defined in Ω whose first-order derivatives are square-integrable in Ω. In fact, the formula

$$2f(\rho) = \inf\{\lambda + \rho^2 \cdot \coth\lambda : \lambda > 0\},$$

which holds for every nonnegative ρ and follows from (1.10), implies either

$$2J(u) \le \lambda \cdot \int n^2 dxdy + \coth\lambda \cdot \int |\nabla u|^2 dxdy$$

for every u and every positive λ or

$$J(u) \le \int n^2 dxdy \times f\left(\sqrt{\frac{\int |\nabla u|^2 dxdy}{\int n^2 dxdy}}\right)$$

for every u. Moreover, an appropriate analysis shows that

$$\sup\left\{\frac{f(\rho_1)-f(\rho_2)}{\rho_1-\rho_2} : 0 \le \rho_1 < \rho_2, \rho_1^2+\rho_2^2 = 2M^2\right\} = f'(M),$$

provided M is positive; hence

$$|J(u_1)-J(u_2)|^2 \le \int |\nabla u_1 - \nabla u_2|^2 dxdy \times \left\{\int n^2 dxdy + \frac{1}{2}\int |\nabla u_1|^2 dxdy + \frac{1}{2}\int |\nabla u_2|^2 dxdy\right\}$$

for every u_1 and u_2.

As a working hypothesis, we propose any member of the domain of J to additionally obey a *boundary condition*, e.g., to take prescribed values on the boundary, $\partial\Omega$, of Ω. (On occasion, ∂ denotes either differentiation or the operation which results in the boundary of a point set.) Formal definitions follow.

(i) $W^{1,2}(\Omega)$ = completion of $C^\infty(\Omega)$ under the norm defined by

$$\|u\|^2_{W^{1,2}(\Omega)} = 4\int_\Omega u^2(x^2+y^2+4)^{-2}\, dxdy + \int_\Omega |\nabla u|^2 dxdy.$$

$W_0^{1,2}(\Omega)$ = closure of $C_0^\infty(\Omega)$ in $W^{1,2}(\Omega)$, i.e., the subset of $W^{1,2}(\Omega)$ consisting of those functions that vanish on $\partial\Omega$ in a generalized sense. (As usual, $C^\infty(\Omega)$ is the set of infinitely differentiable real-valued functions defined in Ω, and $C_0^\infty(\Omega)$ is the subset of $C^\infty(\Omega)$ consisting of those functions that vanish out of a compact subset of Ω.)

(ii) Let j be any given member of $W^{1,2}(\Omega)$; define

$$\text{domain of } J = j + W_0^{1,2}(\Omega), \tag{1.12}$$

i.e., the set of functions u from $W^{1,2}(\Omega)$ such that $u-j$ belongs to $W_0^{1,2}(\Omega)$.

The following assumptions will be made throughout. First, the measure of Ω in Riemannian metric (1.7) is *finite*; i.e.,

$$\int_\Omega n^2 dxdy < \infty. \tag{1.13}$$

Second, Ω is *essentially different* from $\mathbb{R}^2$; i.e.,

$$\text{measure of } (\mathbb{R}^2 \setminus \Omega) > 0. \tag{1.14}$$

Note that Ω is allowed to be either bounded or unbounded. (Relevantly to the present context, Ω may be an *exterior domain,* i.e., an open connected set whose complement is compact.) In the former case, the measure $(x^2+y^2+4)^{-2}dxdy$, appearing

in (i) above, may be virtually replaced by the standard Lebesgue measure $dxdy$; hence $W^{1,2}(\Omega)$ coincides with the collection of functions that are square-integrable in Ω and whose first-order weak derivatives are square-integrable in Ω—a standard Sobolev space. In any case, the measure in question can be thought of as the area element on the two-dimensional unit sphere $\mathbb{S}^2$ parametrized via a stereographic projection; hence $W^{1,2}(\Omega)$ can be identified with a space of standard Sobolev functions defined in an open subset of $\mathbb{S}^2$.

Theorem 2.1 below claims that J does possess a minimum and that the relevant minimizer is unique within the domain specified above.

Since J was born convex, a necessary and sufficient condition for a member of the domain of J to render J a minimum is the Euler–Lagrange equation. J fails to be smoothly differentiable, however. Therefore, the Euler–Lagrange equation of J involves a set-valued subdifferential and must be cast in the form of an inclusion. Details follow.

Let u belong to the domain of J. If φ is any test function, i.e., any member of $W_0^{1,2}(\Omega)$, we have

$$J(u+\varphi)-J(u)=\int_{\{(x,y):\nabla u(x,y)\neq 0\}}\left[f\left(\frac{|\nabla u+\nabla\varphi|}{n}\right)-f\left(\frac{|\nabla u|}{n}\right)\right]n^2dxdy$$
$$+\int_{\{(x,y):\nabla u(x,y)=0\}}f\left(\frac{|\nabla\varphi|}{n}\right)n^2dxdy;$$

moreover,

$$t^{-1}\int_{\{(x,y):\nabla u(x,y)\neq 0\}}\left[f\left(\frac{|\nabla u+t\nabla\varphi|}{n}\right)-f\left(\frac{|\nabla u|}{n}\right)\right]n^2dxdy$$
$$\to\int_{\{(x,y):\nabla u(x,y)\neq 0\}}\frac{n}{|\nabla u|}f'\left(\frac{|\nabla u|}{n}\right)(\nabla u,\nabla\varphi)\,dxdy$$

as t approaches 0, and

$$t^{-1}\int_{\{(x,y):\nabla u(x,y)=0\}}f\left(\frac{t|\nabla\varphi|}{n}\right)n^2dxdy$$
$$\to f'(0)\cdot\int_{\{(x,y):\nabla u(x,y)=0\}}|\nabla\varphi|\;n\;dxdy$$

as t approaches 0 through *positive* values. Therefore,

$$\lim_{t\downarrow 0}\left[J(u+t\varphi)-J(u)\right]/t,$$

the *one-sided directional derivative* of J at u with respect to φ, equals

$$\int_{\{(x,y):\nabla u(x,y)\neq 0\}}\sqrt{1+n^2|\nabla u|^{-2}}\,(\nabla u,\nabla\varphi)\,dxdy+\int_{\{(x,y):\nabla u(x,y)=0\}}|\nabla\varphi|\;n\;dxdy.$$

Recall that the *subdifferential* of J, ∂J, may be characterized thusly: (i) $\partial J(u)$ is a *convex set* of distributions; (ii) a distribution T belongs to $\partial J(u)$ if and only if the directional derivative of J at u with respect to φ is greater than or equals $T(\varphi)$ for

every test function φ. Consequently, $\partial J(u)$ is the collection of those distributions T satisfying

$$\int_{\{(x,y):\nabla u(x,y)\neq 0\}} \sqrt{1+n^2|\nabla u|^{-2}}\,(\nabla u,\nabla\varphi)\,dxdy + \int_{\{(x,y):\nabla u(x,y)=0\}} |\nabla\varphi|\ n\ dxdy \geq T(\varphi)$$

for every test function, φ. Such a formula implies that $\partial J(u)\neq\emptyset$, i.e., that J is *everywhere subdifferentiable*, and, moreover, that any member T of $\partial J(u)$ obeys

$$T=-\mathrm{div}\{\sqrt{1+n^2|\nabla u|^{-2}}\nabla u\}$$

in any open set $\mathcal{O}$ contained in Ω and essentially contained in

$$\{(x,y)\in\Omega;\nabla u(x,y)\neq 0\},$$

i.e., satisfying

$$\text{measure of } \mathcal{O}\cap\{(x,y)\in\Omega:\nabla u(x,y)=0\}=0.$$

We see, in particular, that J is *differentiable* at u if the set of the critical points of u has measure *zero; J fails* to be differentiable at u if the set of the critical points of u has a *positive* measure.

The analysis provided may be summarized in this way. The appropriate *Euler–Lagrange equation* of J reads

$$\partial J(u)\ni 0,$$

an *inclusion* that implies the following: (1.4) holds in the sense of distributions in any open subset of Ω which is essentially contained in $\{(x,y)\in\Omega:\nabla u(x,y)\neq 0\}$.

In other words, a solution u to the Euler–Lagrange equation of J solves a *free boundary problem* for (1.4), the relevant *free boundary* being

$$\Omega\cap\partial\{(x,y)\in\Omega:\nabla u(x,y)\neq 0\}.$$

(Let a manifold $\mathfrak{M}$, a class of nice functions defined in $\mathfrak{M}$, and a differential equation be given. Suppose a member u of the given function class and a subset $\mathfrak{N}$ of $\mathfrak{M}$ are sought such that (i) u solves the given equation in any open subset of $\mathfrak{N}$ or in any open set which is essentially contained in $\mathfrak{N}$; (ii) u obeys special conditions either on $\partial\mathfrak{N}\cap\mathfrak{M}$ or out of $\mathfrak{N}$. It is usual to say that a free boundary problem is in hand. $\partial\mathfrak{N}\cap\mathfrak{M}$, the boundary of $\mathfrak{N}$ relative to $\mathfrak{M}$, is called the free boundary. [16] and [19] are exhaustive references on this matter.)

What is the geometry and the physical meaning of these free boundaries? The results recorded in the present paper, though not equal to a full proof, give evidence to the following statements. The free boundaries in question (i) either are empty or are genuine *curves*—rather than collections of isolated points; and (ii) separate regions where evanescent waves develop from regions where geometrical optics prevails—hence coincide with *caustics.* (Recall that the envelopes of rays are nicknamed caustics, and thus caustics are precisely the contours near and beyond which geometrical optics break down.)

Samples of free boundaries, which affect solutions to (1.4), appear in [26, section 2.4] or can be detected in Figure 1.1.

1.4. Summary of results. We have sketched an existence result that is a key to our investigations; i.e., the *minimizer u of an apposite functional both takes prescribed boundary values and solves a free boundary value problem for* (1.4). The main issues of the present paper, which are detailed in section 2, can be summarized as follows.

Suppose n is differentiable and its first-order derivatives belong to $L^2_{\mathrm{loc}}(\Omega)$. (As usual, $L^2(\Omega)$ is the space of the real-valued functions that are square-integrable in Ω, and $L^2_{\mathrm{loc}}(\Omega)$ is the space of functions φ such that $\varphi \cdot \psi$ belongs to $L^2(\Omega)$ for every ψ from $C_0^\infty(\Omega)$. Occasionally, we will need to replace 2 by some exponent p larger than or equal to 1.)

(i) u is locally twice differentiable in a suitable generalized sense and obeys (1.5) *in the whole* of domain Ω.

(ii) u is a *viscosity solution* to (1.5).

We loosely imitate ideas from [7], [8], [12], and [13, Chapter 10] and mean the following: u_ε approaches u in an appropriate topology as ε approaches zero. Here ε is a strictly positive constant parameter, and u_ε is the twice differentiable real-valued function that obeys a *restored version* of (1.5) and takes the relevant boundary values. Such a version results from adding the extra term

$$\varepsilon \cdot n^2 \left(n^2 + |\nabla u|^2\right) \cdot \Delta u$$

to the l.h.s. of (1.5), i.e., reads

$$\begin{aligned} &\varepsilon \cdot n^2 \left(n^2 + |\nabla u|^2\right) \cdot \Delta u \\ &+ \left\{|\nabla u|^4 + n^2 u_y^2\right\} \cdot u_{xx} - 2n^2 u_x u_y \cdot u_{xy} + \left\{|\nabla u|^4 + n^2 u_x^2\right\} \cdot u_{yy} \\ &+ n|\nabla u|^2 (\nabla n \cdot \nabla u) = 0. \end{aligned} \tag{1.15}$$

Observe that (1.15) is *uniformly elliptic* and that its leading part balances the first-order terms properly; in other words, the injection of viscosity cures degeneracy. In fact, if a_{11}, a_{12}, and a_{22} denote the coefficients of u_{xx}, u_{xy}, and u_{yy} in (1.15) and ρ and ω are defined by

$$|\nabla u| = n\rho, \quad u_x : \cos\omega = u_y : \sin\omega,$$

then

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$$

$$= n^4 \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix} \begin{bmatrix} \rho^4 + \varepsilon(1+\rho^2) & 0 \\ 0 & (1+\rho^2)(\varepsilon+\rho^2) \end{bmatrix} \begin{bmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{bmatrix}. \tag{1.16}$$

Therefore, the eigenvalues of $[a_{ij}]$ obey

$$\frac{\text{smaller eigenvalue}}{\text{larger eigenvalue}} \geq \sqrt{\varepsilon} \cdot (2 + \sqrt{\varepsilon})(1 + \sqrt{\varepsilon})^{-2},$$

and we have

$$\frac{|\text{first-order term}|}{\text{larger eigenvalue}} \leq (1 + \sqrt{\varepsilon})^{-2} \times \frac{|\nabla n|}{n} \times \text{the first power of } |\nabla u|.$$

Viscosity solutions are focused on in section 5, where we show that (i) a viscosity solution to (1.5) is uniquely determined by its boundary values; (ii) a smooth solution to the same equation need not do the same—therefore, a smooth solution to (1.5) need not be a viscosity solution.

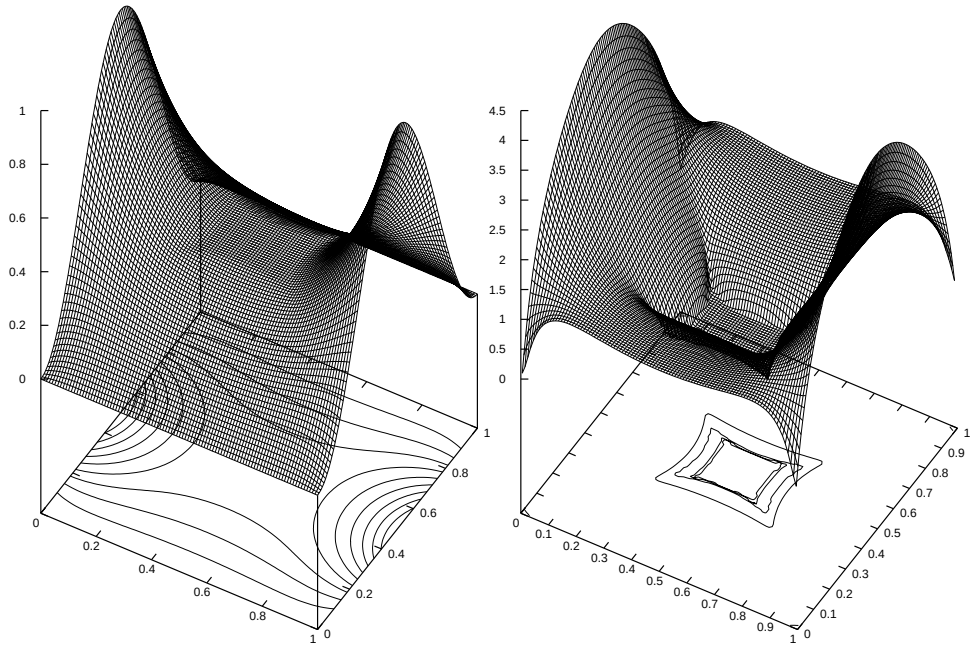


FIG. 1.1. *Typical plots of u and $|\nabla u|$. Here u is a viscosity solution to* (1.5).

1.5. Future developments. Viscosity solutions to (1.5) can be computed efficiently either by finite difference methods or by finite element methods. Details and relevant codes will appear elsewhere.

By way of an example, let u be the viscosity solution that obeys (1.5) in the domain

$$]0,1[\times]0,1[$$

and satisfies the following boundary conditions:

$$\begin{aligned} u(x,0) = u(x,1) &= 0 \qquad \text{if } \ 0 \le x \le 1, \\ u(0,y) = u(1,y) &= [\sin(\pi y)]^2 \ \text{ if } \ 0 \le y \le 1. \end{aligned}$$

Figure 1.1 shows plots of u and $|\nabla u|$, respectively. There, u is approximated by the solution to (1.15) that takes the boundary values in hand, $\varepsilon = 10^{-8}$, finite differences are used, and a 200×200 uniform grid is involved. Note a peculiarity—the solution in question develops caustics, i.e., an inner plateau.

In part three of our work, which will be assembled in a future paper, we will show how the present results, Bäcklund trasformations, and suitable extra ingredients supply solutions to either (1.1) or (1.2) and guarantee their uniqueness.

The referees pointed out that Theorem 9.3 from [9] should be referenced here. Such a theorem claims that if Ω is any open subset of $\mathbb{R}^2$, φ is any Lipschitz continuous map from Ω into $\mathbb{R}^2$, and n is real-valued and continuous, then system (1.2) admits solutions that are Lipschitz continuous in Ω and equal to φ on $\partial\Omega$.

This theorem departs from our point of view for a couple of reasons. First, we are interested in tractable solutions, i.e., smooth enough, unique, and actually computable. Second, we do not address system (1.2) in the present paper. Treating (1.2)

by the present methods cannot be done in few words and deserves further investigation.

2. Main results. Let J be defined by (1.9), (1.10), (1.11), and (1.12). Assume conditions (1.13) and (1.14).

Let ε be a parameter satisfying

$$0 < \varepsilon \le 1/2.$$

Let a real function f_ε be defined by

$$f_\varepsilon(\rho) = \int_0^\rho t \left(\frac{1+t^2}{\varepsilon + t^2} \right)^{\frac{1}{2(1-\varepsilon)}} dt \tag{2.1}$$

for every nonnegative ρ; let a functional J_ε be defined by

$$\begin{aligned} &\text{domain of } J_\varepsilon = \text{domain of } J, \\ &J_\varepsilon(u) = \int_\Omega f_\varepsilon \left(\frac{|\nabla u|}{n} \right) n^2 dxdy. \end{aligned} \tag{2.2}$$

THEOREM 2.1. *Functional J achieves a minimum and has a unique minimizer.*

THEOREM 2.2. (i) *Functional J_ε achieves a minimum and has a unique minimizer.*

(ii) *Let u and u_ε denote the minimizer of J and the minimizer of J_ε, respectively; then u_ε converges to u both in $L^2_{\mathrm{loc}}(\Omega)$ and weakly in $W^{1,2}(\Omega)$ as ε approaches 0.*

THEOREM 2.3. *Suppose n is differentiable and the first-order derivatives of n belong to $L^2_{\mathrm{loc}}(\Omega)$; let u and be u_ε be as above. We make the following assertions:*

(i) *u_ε is twice differentiable in the usual generalized sense, the second-order derivatives of u_ε belong to $L^2_{\mathrm{loc}}(\Omega)$, and u_ε obeys* (1.15).

(ii) *u_ε converges to u uniformly on every compact subset of Ω as ε approaches* 0; *∇u_ε converges to ∇u in $L^p_{\mathrm{loc}}(\Omega) \times L^p_{\mathrm{loc}}(\Omega)$ for every p larger than or equal to* 1.

(iii) *u is twice differentiable in a generalized sense and obeys the inequality*

$$\begin{aligned} &\left\{ \int_{\{(x,y):\mathrm{dist}((x,y),\mathbb{R}^2\setminus K)\ge r\}} \frac{|\nabla u|^2}{n^2+|\nabla u|^2} \left(u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2 \right) \, dxdy \right\}^{\frac{1}{2}} \\ &\qquad \le 6 \left\{ \int_K |\nabla n|^2 \, dxdy \right\}^{\frac{1}{2}} + 2r^{-1} \left\{ \int_K \left(n^2 + |\nabla n|^2 \right) \, dxdy \right\}^{\frac{1}{2}} \end{aligned} \tag{2.3}$$

provided that K is a nice compact subset of Ω and r is a positive number. Moreover, u makes

$$(n^2 + |\nabla u|^2)^{-\frac{3}{2}} \times \{\text{l.h.s. of (1.5)}\}$$

both locally integrable in Ω and equal to 0*; in other words, u obeys* (1.5) *in the whole of Ω.*

3. Proofs of Theorems 2.1 and 2.2.

3.1. An inequality. A proof of Theorem 2.1 relies upon the following lemma.

LEMMA 3.1. *Let Ω obey* (1.9) *and* (1.14), *and let C be any constant such that*

$$C \ge \left\{ \frac{4}{\pi} \int_{\mathbb{R}^2\setminus\Omega} \frac{dxdy}{(x^2+y^2+4)^2} \right\}^{-1}.$$

Then

$$4\int_{\mathbb{R}^2}\varphi^2(x^2+y^2+4)^{-2}\,dxdy \le (C-1)\int_{\mathbb{R}^2}\left(\varphi_x^2+\varphi_y^2\right)\,dxdy$$

provided that φ is smooth enough and real-valued, and

$$\textit{support of } \varphi \subseteq \Omega.$$

Proof. The metric induced by

$$\frac{(dx)^2+(dy)^2}{\left[1+(x^2+y^2)/4\right]^2}$$

makes $\mathbb{R}^2$ a Riemannian manifold $\mathfrak{M}$ that is locally conformal to a unit sphere and enjoys the following properties. First, the Riemannian area element equals

$$\left[1+(x^2+y^2)/4\right]^{-2}\,dxdy.$$

Second, the length of the Riemannian gradient of any smooth scalar field equals

$$\left[1+(x^2+y^2)/4\right]\times \text{length of the Euclidean gradient.}$$

Thus the Riemannian area of $\mathbb{R}^2$ equals 4π, and

$$\text{Riemannian area of } \Omega \le 4\pi(1-1/C);$$

moreover,

$$\int_{\mathbb{R}^2}\varphi^2\left[1+(x^2+y^2)/4\right]^{-2}\,dxdy=\int_{\mathfrak{M}}\varphi^2,$$

and

$$\int_{\mathbb{R}^2}\left(\varphi_x^2+\varphi_y^2\right)\,dxdy=\int_{\mathfrak{M}}|\text{Riemannian gradient of } \varphi|^2.$$

We must show that

$$\int_{\mathfrak{M}}\varphi^2 \le 4(C-1)\int_{\mathfrak{M}}|\text{Riemannian gradient of } \varphi|^2. \tag{3.1}$$

Let μ and φ^* be the distribution function and the decreasing rearrangement of φ, respectively. μ is the map from $[0,\infty[$ into $[0,4\pi]$ such that

$$\mu(t)=\text{Riemannian area of } \{(x,y):|\varphi(x,y)|>t\}$$

for every nonnegative t. φ^* can be defined as the map from $[0,4\pi]$ into $[0,\infty[$ which is right-continuous, decreasing, and equidistributed with φ, i.e., such that

$$\text{length of } \{s\in[0,4\pi]:\varphi^*(s)>t\}=\mu(t)$$

for every nonnegative t.

We have

$$\int_{\mathfrak{M}}\varphi^2=\int_0^{4\pi}\left[\varphi^*(s)\right]^2 ds \tag{3.2}$$

since the very definitions of μ and φ^* ensure that both sides equal $\int_0^\infty t^2[-d\mu(t)]$. On the other hand, a version of an important inequality (see, e.g., [2, section 4]) tells us that φ^* is locally absolutely continuous in $]0, 4\pi[$ and satisfies

$$\int_{\mathfrak{M}} |\text{Riemannian gradient of } \varphi|^2 \geq \int_0^{4\pi} s(4\pi - s) \left[-\frac{d\varphi^*}{ds}(s) \right]^2 \, ds. \tag{3.3}$$

The support of φ^* is an interval whose endpoints are 0 and the Riemannian area of the support of φ. Therefore, our hypotheses yield

$$\text{support of } \varphi^* \subseteq [0, 4\pi(1 - 1/C)].$$

Such an inclusion informs us that φ^* vanishes in a neighborhood of 4π. Thus an integration by parts and a Schwarz inequality give successively

$$\int_0^{4\pi} [\varphi^*(s)]^2 \, ds = 2 \int_0^{4\pi} \varphi^*(s) s \left[-\frac{d\varphi^*}{ds}(s) \right] \, ds$$

and

$$\int_0^{4\pi} [\varphi^*(s)]^2 \, ds \leq 4 \int_0^{4\pi} s^2 \left[-\frac{d\varphi^*}{ds}(s) \right]^2 \, ds.$$

The same inclusion also implies that

$$\int_0^{4\pi} s^2 \left[-\frac{d\varphi^*}{ds}(s) \right]^2 \, ds \leq (C-1) \int_0^{4\pi} s(4\pi - s) \left[-\frac{d\varphi^*}{ds}(s) \right]^2 \, ds.$$

We infer

$$\int_0^{4\pi} [\varphi^*(s)]^2 \, ds \leq 4(C-1) \int_0^{4\pi} s(4\pi - s) \left[-\frac{d\varphi^*}{ds}(s) \right]^2 \, ds. \tag{3.4}$$

Equation (3.2) and inequalities (3.3) and (3.4) result in (3.1). □

3.2. Proof of Theorem 2.1. Uniqueness of the minimizer results from the strict convexity of functional J, while existence follows from the items below via standard arguments of the calculus of variations.

(i) *Boundedness of sublevel sets of J.* The formula

$$f(\rho) = \sup \left\{ \rho \cdot \frac{\lambda}{\sinh \lambda} + \rho^2 \cdot \frac{\sinh(2\lambda) - 2\lambda}{4(\sinh \lambda)^2} : \lambda > 0 \right\},$$

which holds for every nonnegative ρ and follows from (1.10), gives successively

$$J(u) \geq \frac{\lambda}{\sinh \lambda} \cdot \int |\nabla u| \; n \; dxdy + \frac{\sinh(2\lambda) - 2\lambda}{4(\sinh \lambda)^2} \cdot \int |\nabla u|^2 dxdy$$

for every u and every positive λ and either

$$J(u) \geq \frac{\left(\int |\nabla u| n dxdy \right)^2}{\int |\nabla u|^2 dxdy} \times f \left(\frac{\int |\nabla u|^2 dxdy}{\int |\nabla u| n dxdy} \right)$$

or

$$J(u) \geq \int |\nabla u| n dx dy,$$

or else

$$J(u) \geq \frac{1}{2} \int |\nabla u|^2 dx dy \tag{3.5}$$

for every u.

Lemma 3.1 implies that every u from $j + W_0^{1,2}(\Omega)$ obeys

$$\|u\|_{W^{1,2}(\Omega)} \leq (1 + \sqrt{C}) \cdot \|j\|_{W^{1,2}(\Omega)} + \sqrt{C} \cdot \left\{ \int |\nabla u|^2 dx dy \right\}^{1/2}. \tag{3.6}$$

Inequality (3.5) tells us that J is *coercive.* Inequalities (3.5) and (3.6) imply that the sublevel sets of J, i.e., the function classes

$$\{u \in \text{domain of } J : J(u) \leq \text{Constant}\},$$

are all *bounded* in the metric of $W^{1,2}(\Omega)$.

(ii) *Compactness.* The classical Riesz compactness theorem or an oversimplified version of the Rellich–Kondrachov theorem (see, e.g., [1, Chapter V] or [34, section 2.5]) ensures that any sequence which is bounded in $W^{1,2}(\Omega)$ contains some subsequence which converges in $L^2_{\text{loc}}(\Omega)$. The structure of the appropriate dual space (see, e.g., [1, Chapter III] or [34, section 4.3]) ensures that any sequence which is bounded in $W^{1,2}(\Omega)$ and converges in $L^2_{\text{loc}}(\Omega)$ does converge in the weak topology of $W^{1,2}(\Omega)$.

(iii) *Lower semicontinuity of J.* The real function g defined by

$$\begin{aligned} g(\rho) &= 0 && \text{if} \quad 0 \leq \rho \leq 1, \\ &= \tfrac{1}{2}[\rho\sqrt{\rho^2 - 1} - \log(\rho + \sqrt{\rho^2 - 1})] && \text{if} \quad \rho > 1 \end{aligned}$$

is the *Young conjugate* of f, i.e., obeys

$$f(\rho) = \sup\{\rho \cdot \lambda - g(\lambda) : \lambda \geq 0\}$$

for every nonnegative ρ. Therefore, either an inspection or a theorem from [29] gives

$$J(u) = \sup \left\{ \int (\nabla u, \varphi)\, dx dy - \int g\left(\frac{|\varphi|}{n}\right) n^2 dx dy : \varphi \in L^2(\Omega) \times L^2(\Omega) \right\} \tag{3.7}$$

for every u.

Since $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$, the former can replace the latter in the preceding formula. Hence an integration by parts gives

$$J(u) = \sup \left\{ -\int u \cdot \operatorname{div} \varphi \; dx dy - \int g\left(\frac{|\varphi|}{n}\right) n^2 dx dy : \varphi \in C_0^\infty(\Omega) \times C_0^\infty(\Omega) \right\} \tag{3.8}$$

for every u.

The supremum of a family of continuous functionals is lower semicontinuous. Thus (3.7) and (3.8) imply that J is *lower semicontinuous* with respect to both the weak topology of $W^{1,2}(\Omega)$ and the topology of $L^2_{\text{loc}}(\Omega)$. □

3.3. Proof of Theorem 2.2. Proposition (i) is a replica of Theorem 2.1 and can be demonstrated similarly. Ad hoc ingredients, such as the convexity and the coerciveness of functional J_ε, are provided by (2.1) and (2.2) and by propositions (i), (ii), and (iii) of Lemma A.1.

Proposition (ii) is straightforward. Since (1.11) and (2.2) give

$$|J(\varphi) - J_\varepsilon(\varphi)| \le \int_\Omega n^2 \, dxdy \times \sup\{|f(\rho) - f_\varepsilon(\rho)| : 0 \le \rho < \infty\}$$

for every φ, proposition (iv) of Lemma A.1 implies that

$$\sup\left\{|J(\varphi) - J_\varepsilon(\varphi)| : \varphi \in W^{1,2}(\Omega)\right\} = O(\sqrt{\varepsilon}); \tag{3.9}$$

that is, J_ε converges *uniformly* to J as ε approaches 0.

On the other hand,

$$0 \le J(u_\varepsilon) - \min J \le 2 \cdot \sup\left\{|J(\varphi) - J_\varepsilon(\varphi)| : \varphi \in W^{1,2}(\Omega)\right\}. \tag{3.10}$$

Formulas (3.9) and (3.10) imply that

$$\lim_{\varepsilon \to 0} J(u_\varepsilon) = \min J;$$

therefore,

$$\{u_{\varepsilon_k}\}_{k=1,2,3,\ldots}$$

is a *minimizing sequence* relative to functional J whenever $\{\varepsilon_k\}_{k=1,2,3,\ldots}$ obeys $0 < \varepsilon_k \le 1/2$ for every k and

$$\lim_{k \to \infty} \varepsilon_k = 0.$$

Suppose, by contradiction, that u_ε fails to approach u either in $L^2_{\mathrm{loc}}(\Omega)$ or in the weak topology of $W^{1,2}(\Omega)$ as ε approaches 0. Then a neighborhood of u and a sequence $\{\varepsilon_k\}_{k=1,2,3,\ldots}$ exist such that u_{ε_k} is out of this neighborhood and $0 < \varepsilon_k \le 1/(2k)$ for every k.

The analysis made in section 3.2, while proving Theorem 2.1, shows that every minimizing sequence relative to J contains a subsequence which converges to a minimizer of J both in $L^2_{\mathrm{loc}}(\Omega)$ and in the weak topology of $W^{1,2}(\Omega)$.

Therefore, a minimizer of J exists which is out of some neighborhood of u and thus is different from u.

This is impossible because J is strictly convex, and a strictly convex functional cannot have two different minimizers. □

4. Proof of Theorem 2.3.

4.1. Proof of proposition (i) of Theorem 2.3. The proof is patterned on conventional arguments of the calculus of variations and consists of the three items below.

(i) *Euler–Lagrange equation of J_ε—weak form.* Proposition (v) of Lemma A.1 tells us that

$$\mathbb{R}^2 \ni (p, q) \mapsto n^2 \cdot f_\varepsilon(n^{-1} \cdot \sqrt{p^2 + q^2})$$

is twice continuously differentiable. If ρ and ω are defined by

$$p = n\rho \cdot \cos\omega \quad \text{and} \quad q = n\rho \cdot \sin\omega,$$

then the gradient of the above function equals

$$f'_\varepsilon(\rho) \cdot \begin{bmatrix} \cos\omega \\ \sin\omega \end{bmatrix},$$

and its Hessian matrix, H, is given by

$$H = \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix} \begin{bmatrix} f''_\varepsilon(\rho) & 0 \\ 0 & f'_\varepsilon(\rho)/\rho \end{bmatrix} \begin{bmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{bmatrix}. \tag{4.1}$$

Proposition (vi) of Lemma A.1 tells us that the eigenvalues involved obey

$$0 < \text{ eigenvalues } \le \varepsilon^{-\frac{1}{2(1-\varepsilon)}}.$$

Therefore, Taylor's formula gives

$$J_\varepsilon(u+\varphi) - J_\varepsilon(u) = \int_\Omega \frac{n}{|\nabla u|} \cdot f'_\varepsilon\left(\frac{|\nabla u|}{n}\right) \cdot (\nabla u, \nabla\varphi)\ dxdy + \text{ a remainder},$$

$$0 \le 2 \cdot (\text{remainder}) \le \varepsilon^{-\frac{1}{2(1-\varepsilon)}} \int_\Omega |\nabla\varphi|^2\ dxdy$$

provided that u and φ are endowed with square-integrable first-order derivatives. We infer that J_ε is differentiable at every u from its domain, and

$$J'_\varepsilon(u)(\varphi) = \int_\Omega \frac{n}{|\nabla u|} \cdot f'_\varepsilon\left(\frac{|\nabla u|}{n}\right) \cdot (\nabla u, \nabla\varphi)\ dxdy$$

for every φ from $W^{1,2}_0(\Omega)$; in other words,

$$J'_\varepsilon(u) = -\mathrm{div}\left\{ n \cdot f'_\varepsilon\left(\frac{|\nabla u|}{n}\right) \cdot \frac{\nabla u}{|\nabla u|} \right\}$$

in the sense of distributions.

The analysis provided shows that the minimizer of J_ε obeys the equation

$$\mathrm{div}\left\{ n \cdot f'_\varepsilon\left(\frac{|\nabla u|}{n}\right) \cdot \frac{\nabla u}{|\nabla u|} \right\} = 0 \tag{4.2}$$

in the sense of distributions. Thus the *Euler–Lagrange equation* of functional J_ε amounts precisely to (4.2).

(ii) *Extra regularity of extremals.* Now we resort to the hypothesis made on n and claim that, if u is a distributional solution to (4.2) and

$$\nabla u \in L^2_{\mathrm{loc}}(\Omega) \times L^2_{\mathrm{loc}}(\Omega),$$

then u is twice differentiable and its second-order derivatives are in $L^2_{\mathrm{loc}}(\Omega)$.

A proof of such a claim can be outlined in this way.

Let ρ and ω be defined by

$$p = n\rho \cdot \cos\omega \quad \text{and} \quad q = n\rho \cdot \sin\omega,$$

let H be defined as in (4.1), and let either

$$F = [f'_\varepsilon(\rho) - \rho f''_\varepsilon(\rho)] \cdot \frac{\partial n}{\partial x} \cdot \begin{bmatrix} \cos\omega \\ \sin\omega \end{bmatrix}$$

or

$$F = [f'_\varepsilon(\rho) - \rho f''_\varepsilon(\rho)] \cdot \frac{\partial n}{\partial y} \cdot \begin{bmatrix} \cos\omega \\ \sin\omega \end{bmatrix};$$

consider the partial differential equation

$$\operatorname{div}(H \cdot \nabla v) = \operatorname{div} F. \tag{4.3}$$

Proposition (ii) of Lemma 4.1 tells us that certain constants, depending only upon ε, exist such that

$$0 < \text{ Constant } \leq \text{ eigenvalues of } H \leq \text{ Constant},$$

and

$$|F| \leq \text{ Constant} \cdot |\nabla n|.$$

As a consequence, it can be shown that another constant, depending upon ε, exists such that

$$\begin{aligned} &\int_{\{(x,y):\operatorname{dist}((x,y),\mathbb{R}^2\setminus K)\geq r\}} |\nabla v|^2 \, dxdy \\ &\leq \text{ Constant} \cdot \left[\int_K |\nabla n|^2 \, dxdy + r^{-2} \int_K v^2 \, dxdy \right] \end{aligned} \tag{4.4}$$

provided that v is any distributional solution to (4.3), K is a nice compact subset of Ω, and r is a positive number.

Inequality (4.4), which is sometimes referred to as *Caccioppoli's inequality,* plus an appropriate use of finite differences allow one to conclude that, if either

$$v = \partial u/\partial x$$

or

$$v = \partial u/\partial y,$$

then v actually obeys (4.3) in the sense of distributions and

$$\nabla v \in L^2_{\text{loc}}(\Omega) \times L^2_{\text{loc}}(\Omega).$$

Details can be found, e.g., in [17, section 2.1], [22, sections 4.3 and 4.5], [27, seciton 1.10 and 1.11]. The claim is demonstrated.

(iii) *Euler–Lagrange equation of J_ε—strong form.* An appropriate smoothness and appropriate symbols of relevant ingredients having been established, (4.3) can be recast in the following form:

$$\begin{aligned} &\left[f''_\varepsilon(\rho)(\cos\omega)^2 + \frac{f'_\varepsilon(\rho)}{\rho}(\sin\omega)^2 \right] u_{xx} + 2\left[f''_\varepsilon(\rho) - \frac{f'_\varepsilon(\rho)}{\rho} \right] \cos\omega \ \sin\omega \ u_{xy} \\ &+ \left[f''_\varepsilon(\rho)(\sin\omega)^2 + \frac{f'_\varepsilon(\rho)}{\rho}(\cos\omega)^2 \right] u_{yy} + \left[\frac{f'_\varepsilon(\rho)}{\rho} - f''_\varepsilon(\rho) \right] \nabla u \cdot \nabla \log n = 0. \end{aligned} \tag{4.5}$$

As observed in the appendix, (2.1) implies (A.1) and (A.2); these equations give

$$\frac{f'_\varepsilon(\rho)}{\rho} \ : \ \big[(1+\rho^2)(\varepsilon+\rho^2)\big] = \left[\frac{f'_\varepsilon(\rho)}{\rho} - f''_\varepsilon(\rho)\right] \ : \ \rho^2 = (1+\rho^2)^{-\frac{1-2\varepsilon}{2(1-\varepsilon)}}(\varepsilon+\rho^2)^{-\frac{3-2\varepsilon}{2(1-\varepsilon)}}$$

for every nonnegative ρ.

Consequently, (4.5) coincides with (1.15). In other words, (1.15) is another form of the Euler–Lagrange equation for J_ε. □

4.2. Two lemmas. A proof of proposition (ii) of Theorem 2.3 relies upon the following lemmas.

LEMMA 4.1. *Suppose A and B are 2×2 real symmetric matrices. Let A be positive definite, and let*

$$\kappa = \frac{\textit{smaller eigenvalue}}{\textit{larger eigenvalue}},$$

a condition number of A. Then

$$\frac{(\mathrm{tr}AB)^2}{\det A} - 2\cdot\det B \geq \kappa\cdot\mathrm{tr}(B^2). \tag{4.6}$$

(Here tr *and* det *stand for trace and determinant, respectively.)*

Proof. Denote the entries of A and B by a_{ij} and b_{ij}, respectively; let

$$\boldsymbol{M} = \frac{1}{a_{11}a_{22}-a_{12}^2}\begin{bmatrix} a_{11}^2 & \sqrt{2}a_{11}a_{12} & a_{12}^2 \\ \sqrt{2}a_{11}a_{12} & a_{11}a_{22}+a_{12}^2 & \sqrt{2}a_{12}a_{22} \\ a_{12}^2 & \sqrt{2}a_{12}a_{22} & a_{22}^2 \end{bmatrix}$$

and

$$\boldsymbol{m} = \begin{bmatrix} b_{11} \\ \sqrt{2}b_{12} \\ b_{22} \end{bmatrix}.$$

We have

$$\frac{(\mathrm{tr}AB)^2}{\det A} - 2\cdot\det B = (\boldsymbol{M}\boldsymbol{m},\boldsymbol{m}), \quad \mathrm{tr}(B^2) = (\boldsymbol{m},\boldsymbol{m}).$$

An inspection shows that the eigenvalues of $\boldsymbol{M}$ are $1/\kappa, 1, \kappa$. Inequality (4.6) follows. □

LEMMA 4.2. *Let a real-valued function t be defined by*

$$t(\rho) = \tan\left(\frac{1}{2}\arctan\rho\right) \tag{4.7}$$

for every nonnegative ρ, and let a mapping T be defined by

$$T\varphi = t\left(\frac{|\nabla\varphi|}{n}\right)\ \nabla\varphi \tag{4.8}$$

for every φ from a space of sufficiently smooth real-valued functions of x and y. Assume $\nabla(T\varphi)$ stands for the Jacobian matrix of $T\varphi$ and

$$|\nabla(T\varphi)| = \sqrt{\mathrm{tr}\,[(\nabla(T\varphi))(\nabla(T\varphi))^T]},$$

a norm for such a matrix.

(i) *If ρ and ω are defined by $|\nabla\varphi| = n\ \rho$, $\varphi_x : \cos\omega = \varphi_y : \sin\omega$, the following equations hold:*

$$\begin{aligned}\nabla(T\varphi) = &-2\left(\sin\left(\frac{1}{2}\arctan\rho\right)\right)^2 \begin{bmatrix} n_x\cos\omega & n_y\cos\omega \\ n_x\sin\omega & n_y\sin\omega \end{bmatrix} \\ &+ \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2}+\frac{1}{2}(t(\rho))^2 \end{bmatrix} \begin{bmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{bmatrix} \\ &\times \frac{|\nabla\varphi|}{\sqrt{n^2+|\nabla\varphi|^2}} \begin{bmatrix} \varphi_{xx} & \varphi_{xy} \\ \varphi_{xy} & \varphi_{yy} \end{bmatrix}, \end{aligned} \tag{4.9}$$

$$\begin{aligned} t\left(\frac{|\nabla\varphi|}{n}\right) &\begin{bmatrix} \varphi_{xx} & \varphi_{xy} \\ \varphi_{xy} & \varphi_{yy} \end{bmatrix} = (t(\rho))^2 \begin{bmatrix} n_x\cos\omega & n_y\cos\omega \\ n_x\sin\omega & n_y\sin\omega \end{bmatrix} \\ &+ \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix} \begin{bmatrix} \frac{1}{2}+\frac{1}{2}(t(\rho))^2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{bmatrix} \nabla(T\varphi). \end{aligned} \tag{4.10}$$

(ii) *The following inequalities hold:*

$$|\nabla(T\varphi)| \le |\nabla n| + \frac{|\nabla\varphi|}{\sqrt{n^2+|\nabla\varphi|^2}}\sqrt{\varphi_{xx}^2+2\varphi_{xy}^2+\varphi_{yy}^2}, \tag{4.11}$$

$$\frac{|\nabla\varphi|}{2\sqrt{n^2+|\nabla\varphi|^2}}\sqrt{\varphi_{xx}^2+2\varphi_{xy}^2+\varphi_{yy}^2} \le |\nabla n| + |\nabla(T\varphi)|. \tag{4.12}$$

(iii) *If φ_1 and φ_2 are real-valued and sufficiently smooth, then*

$$|T\varphi_1 - T\varphi_2| \le |\nabla\varphi_1 - \nabla\varphi_2| \tag{4.13}$$

and

$$|\nabla\varphi_1 - \nabla\varphi_2| \le |T\varphi_1 - T\varphi_2|^{\frac{1}{2}} \cdot (4n + |T\varphi_1 - T\varphi_2|)^{\frac{1}{2}}. \tag{4.14}$$

Proof. Equation (4.7) provides us with the properties

$$t(\rho) = \frac{\rho}{1+\sqrt{1+\rho^2}},$$

$$\begin{aligned} t(\rho) = \frac{\rho}{2\sqrt{1+\rho^2}}\left[1+(t(\rho))^2\right], &\quad t(\rho) = \frac{\rho}{2}\left[1-(t(\rho))^2\right], \\ 0 \le t(\rho) < 1, &\quad \frac{\rho}{2\sqrt{1+\rho^2}} \le t(\rho) < \frac{\rho}{\sqrt{1+\rho^2}}, \\ \rho^2 t'(\rho) = 2\left(\sin\left(\frac{1}{2}\arctan\rho\right)\right)^2, &\quad (\rho t(\rho))' = \frac{\rho}{\sqrt{1+\rho^2}}, \end{aligned} \tag{4.15}$$

which hold for every nonnegative ρ and play a role below.

Differentiating both sides of (4.8) gives

$$\nabla(T\varphi) = -\rho^2 t'(\rho) \begin{bmatrix} n_x \cos\omega & n_y \cos\omega \\ n_x \sin\omega & n_y \sin\omega \end{bmatrix}$$

$$+ \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix} \begin{bmatrix} (\rho t(\rho))' & 0 \\ 0 & t(\rho) \end{bmatrix} \begin{bmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{bmatrix} \begin{bmatrix} \varphi_{xx} & \varphi_{xy} \\ \varphi_{xy} & \varphi_{yy} \end{bmatrix}.$$

Equation (4.9) follows because of equations that appear in (4.15).

Inequalities (4.11) and (4.12) are easily derived from (4.9) and (4.10), respectively, via some matrix algebra and inequalities that appear in (4.15).

Suppose $\nabla\varphi_1 \neq \nabla\varphi_2$. Define ρ_1 and ρ_2 by

$$|\nabla\varphi_1| = n\ \rho_1 \quad \text{and} \quad |\nabla\varphi_2| = n\ \rho_2,$$

respectively; let θ be the angle between $\nabla\varphi_1$ and $\nabla\varphi_2$, i.e., be such that

$$0 \le \theta \le \pi, \quad |\nabla\varphi_1||\nabla\varphi_2|\ \cos\theta = (\nabla\varphi_1, \nabla\varphi_2).$$

Equation (4.8) gives successively

$$\frac{|T\varphi_1 - T\varphi_2|^2}{|\nabla\varphi_1 - \nabla\varphi_2|^2} = \frac{(\rho_1 t(\rho_1))^2 + (\rho_2 t(\rho_2))^2 - 2\rho_1\rho_2 t(\rho_1)t(\rho_2)\ \cos\theta}{\rho_1^2 + \rho_2^2 - 2\rho_1\rho_2\ \cos\theta}$$

and

$$\frac{\partial}{\partial\theta}\frac{|T\varphi_1 - T\varphi_2|^2}{|\nabla\varphi_1 - \nabla\varphi_2|^2} = -2\rho_1\rho_2 \frac{(t(\rho_1) - t(\rho_2))(\rho_1^2 t(\rho_1) - \rho_2^2 t(\rho_2))}{(\rho_1^2 + \rho_2^2 - 2\rho_1\rho_2\ \cos\theta)^2}\ \sin\theta.$$

We have either $t(\rho_1) \le t(\rho_2)$ and $\rho_1^2 t(\rho_1) \le \rho_2^2 t(\rho_2)$ or $t(\rho_1) > t(\rho_2)$ and $\rho_1^2 t(\rho_1) > \rho_2^2 t(\rho_2)$ since both (4.7) and equations in (4.15) show that t is increasing. We infer successively that

$$\frac{\partial}{\partial\theta}\frac{|T\varphi_1 - T\varphi_2|^2}{|\nabla\varphi_1 - \nabla\varphi_2|^2} \le 0$$

and

$$\frac{\rho_1 t(\rho_1) + \rho_2 t(\rho_2)}{\rho_1 + \rho_2} \le \frac{|T\varphi_1 - T\varphi_2|}{|\nabla\varphi_1 - \nabla\varphi_2|} \le \frac{\rho_1 t(\rho_1) - \rho_2 t(\rho_2)}{\rho_1 - \rho_2}.$$

On the other hand, we have

$$t\left(\frac{\rho_1 + \rho_2}{2}\right) \le \frac{\rho_1 t(\rho_1) + \rho_2 t(\rho_2)}{\rho_1 + \rho_2} \quad \text{and} \quad \frac{\rho_1 t(\rho_1) - \rho_2 t(\rho_2)}{\rho_1 - \rho_2} \le 1$$

since equations in (4.15) show that $0 \le \rho \to \rho t(\rho)$ is convex and contractive. Therefore,

$$t\left(\frac{|\nabla\varphi_1| + |\nabla\varphi_2|}{2n}\right) \le \frac{|T\varphi_1 - T\varphi_2|}{|\nabla\varphi_1 - \nabla\varphi_2|} \le 1.$$

We conclude with (4.13) and the inequality

$$t\left(\frac{|\nabla\varphi_1 - \nabla\varphi_2|}{2n}\right)|\nabla\varphi_1 - \nabla\varphi_2| \le |T\varphi_1 - T\varphi_2|,$$

which leads to (4.14) via algebraic manipulations. □

4.3. Proof of proposition (ii) of Theorem 2.3. Suppose K is a compact subset of Ω whose interior is not empty and whose boundary is sufficiently smooth; let $r > 0$, and define

$$\mathcal{K}(r) = \{(x, y) : \mathrm{dist}((x, y), \mathbb{R}^2 \setminus K) \geq r\}.$$

Let T be as in Lemma 4.2.

The following bounds hold.

Bound 1.

$$\int_{\mathcal{K}(r)} \frac{|\nabla u_\varepsilon|^2}{n^2 + |\nabla u_\varepsilon|^2} \left[\left(\frac{\partial^2 u_\varepsilon}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 u_\varepsilon}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 u_\varepsilon}{\partial y^2} \right)^2 \right] dxdy$$

$$\leq \int_K |\nabla n|^2 \, dxdy + r^{-2} \int_K \left(n^2 + |\nabla u_\varepsilon|^2 \right) \, dxdy. \tag{4.16}$$

Bound 2.

$$\left\{ \int_{\mathcal{K}(r)} |\nabla (T u_\varepsilon)|^2 dxdy \right\}^{\frac{1}{2}}$$

$$\leq 2 \left\{ \int_K |\nabla n|^2 \, dxdy \right\}^{\frac{1}{2}} + r^{-1} \left\{ \int_K \left(n^2 + |\nabla u_\varepsilon|^2 \right) \, dxdy \right\}^{\frac{1}{2}}, \tag{4.17}$$

$$\int_K |T u_\varepsilon|^2 \, dxdy \leq \int_K |\nabla u_\varepsilon|^2 \, dxdy. \tag{4.18}$$

Bound 3. If $p \geq 1$, then

$$\left\{ \int_K |\nabla u_{\varepsilon'} - \nabla u_{\varepsilon''}|^p \, dxdy \right\}^2$$

$$\leq \int_K |T u_{\varepsilon'} - T u_{\varepsilon''}|^p \, dxdy \times \int_K (4n + |T u_{\varepsilon'} - T u_{\varepsilon''}|)^p \, dxdy. \tag{4.19}$$

Bound 4.

$$\int_\Omega |\nabla u_\varepsilon|^2 \, dxdy \leq \text{Constant independent of } \varepsilon. \tag{4.20}$$

Proof of Bound 1. For notational convenience, we temporarily drop the subscript ε and denote u_ε by u in short.

We have shown in proposition (i) of Theorem 2.3 that such a u obeys (1.15). Equation (1.15) implies

$$|a_{11} u_{xx} + 2 a_{12} u_{xy} + a_{22} u_{yy}| \leq n \cdot |\nabla u|^3 \cdot |\nabla n|,$$

where a_{11}, a_{12}, and a_{22} are given by (1.16). Equation (1.16) tells us that, in addition to the inequalities appearing in section 1.4, $[a_{ij}]$ satisfies

$$\frac{\text{smaller eigenvalue}}{\text{larger eigenvalue}} \geq \frac{|\nabla u|^2}{n^2 + |\nabla u|^2}$$

and

$$a_{11}a_{22} - a_{12}^2 \geq n^2 \cdot |\nabla u|^6.$$

Therefore, Lemma 4.1 gives

$$\frac{|\nabla u|^2}{n^2 + |\nabla u|^2} \left(u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2\right) \leq 2\left(u_{xy}^2 - u_{xx}u_{yy}\right) + |\nabla n|^2, \tag{4.21}$$

an instance of what is often called *Bernstein's inequality*—see, e.g., [32].

An inspection shows that $u_{xx}u_{yy} - u_{xy}^2$, the Hessian determinant of u, obeys

$$2\left(u_{xy}^2 - u_{xx}u_{yy}\right) = \operatorname{div} \begin{bmatrix} u_{yy} & -u_{xy} \\ -u_{xy} & u_{xx} \end{bmatrix} \cdot \nabla u; \tag{4.22}$$

equivalently,

$$2\left(u_{xy}^2 - u_{xx}u_{yy}\right) dx \wedge dy = d \begin{vmatrix} u_x & u_y \\ du_x & du_y \end{vmatrix}. \tag{4.23}$$

Inequality (4.21) and either (4.22) or (4.23) give

$$\begin{aligned} &\int_{\mathcal{K}(r)} \frac{|\nabla u|^2}{n^2 + |\nabla u|^2} \left(u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2\right) \, dxdy \\ &\quad \leq \left\{ \int_{\partial\mathcal{K}(r)} \left(n^2 + |\nabla u|^2\right) \sqrt{(dx)^2 + (dy)^2} \right\}^{\frac{1}{2}} \\ &\quad \times \left\{ \int_{\partial\mathcal{K}(r)} \frac{|\nabla u|^2}{n^2 + |\nabla u|^2} \left(u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2\right) \, \sqrt{(dx)^2 + (dy)^2} \right\}^{\frac{1}{2}} \\ &\quad + \int_{\mathcal{K}(r)} |\nabla n|^2 \, dxdy \end{aligned} \tag{4.24}$$

via the Gauss–Green formulas and the Cauchy–Schwarz inequality.

If we define two real-valued functions φ and ψ by

$$\varphi(r) = \int_{\mathcal{K}(r)} \frac{|\nabla u|^2}{n^2 + |\nabla u|^2} \left(u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2\right) \, dxdy - \int_K |\nabla n|^2 \, dxdy$$

and

$$\psi(r) = \int_{\mathcal{K}(r)} \left(n^2 + |\nabla u|^2\right) \, dxdy,$$

then a version of the coarea formula (see, e.g., [34, section 2.7] and the equation

$$|\nabla \operatorname{dist}((x,y), \mathbb{R}^2 \setminus K)| = 1 \text{ for almost every } (x,y) \in K$$

yield

$$-\varphi'(r) = \int_{\partial\mathcal{K}(r)} \frac{|\nabla u|^2}{n^2 + |\nabla u|^2} \left(u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2\right) \, \sqrt{(dx)^2 + (dy)^2}$$

and

$$-\psi'(r) = \int_{\partial\mathcal{K}(r)} \left(n^2 + |\nabla u|^2\right) \sqrt{(dx)^2 + (dy)^2}$$

for almost every positive r. Thus (4.24) yields

$$\varphi(r) \le \sqrt{[-\varphi'(r)]\,[-\psi'(r)]} \tag{4.25}$$

for almost every positive r.

As is easy to see, (4.25) implies

$$\text{positive part of } \varphi(r) \le \left\{ \int_0^r \frac{dt}{[-\psi'(t)]} \right\}^{-1}$$

for every positive r. Since

$$r^2 \le \psi(0) \int_0^r \frac{dt}{[-\psi'(t)]},$$

we conclude that

$$\text{positive part of } \varphi(r) \le r^{-2}\psi(0) \tag{4.26}$$

for every positive r.

The inequality

$$\begin{aligned}\int_{\mathcal{K}(r)} \frac{|\nabla u|^2}{n^2 + |\nabla u|^2} \left[\left(\frac{\partial^2 u}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 u}{\partial y^2} \right)^2 \right] dxdy \\ \le \int_K |\nabla n|^2 \, dxdy + r^{-2} \int_K \left(n^2 + |\nabla u|^2\right) dxdy\end{aligned}$$

follows from (4.26). Bound 1 is demonstrated.

Proof of Bound 2. Inequality (4.17) follows from Bound 1 and proposition (ii) of Lemma 4.2. Inequality (4.18) follows from proposition (iii) of Lemma 4.2.

Proof of Bound 3. Such a bound follows from proposition (iii) of Lemma 4.2 and the Cauchy–Schwarz inequality.

Proof of Bound 4. The inequalities

$$\frac{\rho^2}{2} + \frac{1}{4}\log(1 + 2\rho^2) \le f_\varepsilon(\rho) \le f(\rho),$$

which hold for every nonnegative ρ and appear in Lemma A.1, tell us that

$$J_\varepsilon \le J$$

and

$$\frac{1}{2} \int_\Omega |\nabla u_\varepsilon|^2 dxdy \le J_\varepsilon(u_\varepsilon).$$

Since

$$J_\varepsilon(u_\varepsilon) = \min J_\varepsilon,$$

we obtain the inequality

$$\frac{1}{2}\int_\Omega |\nabla u_\varepsilon|^2 dxdy \le \min J,$$

which proves (4.20).

Bound 2, Bound 4, and the Rellich–Kondrachov compactness theorem (see, e.g., [1, Chapter VI] or [34, section 2.5]) ensure that a sequence $\{\varepsilon_k\}_{k=1,2,3,\dots}$ exists such that $0 < \varepsilon_k \le 1/2$ for every k, $\varepsilon_k \to 0$ as $k \to \infty$, and

$$\{Tu_{\varepsilon_k}\}_{k=1,2,3,\dots} \text{ converges in } L^p_{\mathrm{loc}}(\Omega) \times L^p_{\mathrm{loc}}(\Omega)$$

for every p larger than or equal to 1. Consequently, Bound 3 ensures that

$$\{\nabla u_{\varepsilon_k}\}_{k=1,2,3,\dots} \text{ converges in } L^p_{\mathrm{loc}}(\Omega) \times L^p_{\mathrm{loc}}(\Omega)$$

for every p larger than or equal to 1. We now profit by a Sobolev inequality (see, e.g., [1, Chapter V] or [34, section 2.4] and infer that if K is any compact subset of Ω and

$$m_k = (\text{measure of } K)^{-1} \cdot \int_K u_{\varepsilon_k} dxdy,$$

then

$$\{u_{\varepsilon_k} - m_k\}_{k=1,2,3,\dots}$$

converges uniformly in K.

Applying proposition (ii) of Theorem 2.2 leads to the conclusion. □

4.4. Proof of proposition (iii) of Theorem 2.3. Proposition (ii) of Theorem 2.3, proposition (iii) of Lemma 4.2, and Bounds 2 and 4 (appearing in the preceding subsection) show that

$$\{\nabla(Tu_\varepsilon)\}_{k=1,2,3,\dots} \text{ converges weakly in } \left(L^2_{\mathrm{loc}}(\Omega)\right)^4 \tag{4.27}$$

as ε approaches 0.

Having (4.27) in hand, we are in a position to resume a former notation, u, and to establish the ultimate properties of u.

(i) *Second-order derivatives.* Previous ingredients, which include proposition (ii) of Theorem 2.3, Bound 2, and (4.27), guarantee that Tu is differentiable and obeys

$$\begin{aligned} &\left\{\int_{\{(x,y):\mathrm{dist}((x,y),\mathbb{R}^2\setminus K)\ge r\}} |\nabla Tu|^2 dxdy\right\}^{\frac{1}{2}} \\ &\le 2\left\{\int_K |\nabla n|^2 dxdy\right\}^{\frac{1}{2}} + r^{-1}\left\{\int_K \left(n^2 + |\nabla u|^2\right) dxdy\right\}^{\frac{1}{2}} \end{aligned} \tag{4.28}$$

if K is a nice compact subset of Ω and $r > 0$. Propositions (i) and (ii) of Lemma 4.2 and inequality (4.28) show that u is twice differentiable and obeys (2.3).

(ii) *Differential equation.* The underlying idea is recasting both (1.5) and (1.15) in a form in which second-order derivatives of u are replaced by the entries of $\nabla(Tu)$ and then letting ε approach zero. Details follow.

Combining proposition (i) of Lemma 4.2 and the identity

$$|\nabla\varphi|^2\cdot\nabla\left\{n\cdot f_\varepsilon'\left(\frac{|\nabla\varphi|}{n}\right)\cdot\frac{\nabla\varphi}{|\nabla\varphi|}\right\}=n\rho^2[f_\varepsilon'(\rho)/\rho-f_\varepsilon''(\rho)]\begin{bmatrix}n_x\varphi_x & n_y\varphi_x\\ n_x\varphi_y & n_y\varphi_y\end{bmatrix}$$
$$+\begin{bmatrix}\varphi_x & -\varphi_y\\ \varphi_y & \varphi_x\end{bmatrix}\begin{bmatrix}f_\varepsilon''(\rho) & 0\\ 0 & f_\varepsilon'(\rho)/\rho\end{bmatrix}\begin{bmatrix}\varphi_x & \varphi_y\\ -\varphi_y & \varphi_x\end{bmatrix}\begin{bmatrix}\varphi_{xx} & \varphi_{xy}\\ \varphi_{xy} & \varphi_{yy}\end{bmatrix}$$

results in

$$\frac{|\nabla\varphi|^3}{\sqrt{n^2+|\nabla\varphi|^2}}\cdot\nabla\left\{n\cdot f_\varepsilon'\left(\frac{|\nabla\varphi|}{n}\right)\cdot\frac{\nabla\varphi}{|\nabla\varphi|}\right\}$$
$$=n\rho^2t(\rho)\left\{\frac{2f_\varepsilon'(\rho)/\rho}{1+(t(\rho))^2}-f_\varepsilon''(\rho)\right\}\begin{bmatrix}n_x\varphi_x & n_y\varphi_x\\ n_x\varphi_y & n_y\varphi_y\end{bmatrix}$$
$$+\begin{bmatrix}\varphi_x & -\varphi_y\\ \varphi_y & \varphi_x\end{bmatrix}\begin{bmatrix}f_\varepsilon''(\rho) & 0\\ 0 & \frac{2f_\varepsilon'(\rho)/\rho}{1+(t(\rho))^2}\end{bmatrix}\begin{bmatrix}\varphi_x & \varphi_y\\ -\varphi_y & \varphi_x\end{bmatrix}\nabla(T\varphi);\tag{4.29}$$

here φ stands for any sufficiently smooth real-valued function, $\rho=|\nabla\varphi|:n$, and t is given by (4.7).

As observed in the proof of Lemma A.1, (2.1) implies

$$\frac{\rho f_\varepsilon''(\rho)}{f_\varepsilon'(\rho)}=1-\frac{\rho^2}{(1+\rho^2)(\varepsilon+\rho^2)}$$

for every nonnegative ρ. Therefore,

$$0<\rho\cdot\left[\frac{\rho f_\varepsilon''(\rho)}{f_\varepsilon'(\rho)}-\frac{\rho^2}{1+\rho^2}\right]\le\frac{\sqrt{\varepsilon}}{2}\tag{4.30}$$

for every nonnegative ρ. In other words, $\rho^2f_\varepsilon''(\rho)/f_\varepsilon'(\rho)$ converges to $\rho^3/(1+\rho^2)$ *uniformly* with respect to ρ as ε approaches 0.

Mimicking the proof of proposition (iii) of Lemma 4.2 shows that

$$\left|\frac{n}{\sqrt{n^2+|\nabla u_\varepsilon|^2}}\,\nabla u_\varepsilon-\frac{n}{\sqrt{n^2+|\nabla u|^2}}\,\nabla u\right|\le|\nabla u_\varepsilon-\nabla u|.\tag{4.31}$$

Proposition (ii) of Theorem 2.3, (4.27), (4.29), and inequalities (4.30) and (4.31) enable us to conclude that

$$\frac{n^{-4}|\nabla u_\varepsilon|^4}{[1+n^{-2}|\nabla u_\varepsilon|^2]^{3/2}}\cdot\frac{|\nabla u_\varepsilon|}{f_\varepsilon'\left(n^{-1}|\nabla u_\varepsilon|\right)}\cdot\operatorname{div}\left\{n\cdot f_\varepsilon'\left(\frac{|\nabla u_\varepsilon|}{n}\right)\cdot\frac{\nabla u_\varepsilon}{|\nabla u_\varepsilon|}\right\}\tag{4.32}$$

approaches

$$\frac{|\nabla u|}{n^2+|\nabla u|^2}\cdot\operatorname{tr}\left\{n\rho^2t(\rho)\left[\frac{2}{1+(t(\rho))^2}-\frac{\rho^2}{1+\rho^2}\right]\begin{bmatrix}n_xu_x & n_yu_x\\ n_xu_y & n_yu_y\end{bmatrix}\right.$$
$$\left.+\begin{bmatrix}u_x & -u_y\\ u_y & u_x\end{bmatrix}\begin{bmatrix}\frac{\rho^2}{1+\rho^2} & 0\\ 0 & \frac{2}{1+(t(\rho))^2}\end{bmatrix}\begin{bmatrix}u_x & u_y\\ -u_y & u_x\end{bmatrix}\nabla(Tu)\right\}\tag{4.33}$$

in $L^1_{\mathrm{loc}}(\Omega)$ as ε approaches 0.

If expression (4.33) is named A and

$$U = \text{l.h.s. of (1.5)},$$

then (4.7) and (4.8) cause the following equation to hold:

$$A = |\nabla u|^2 \times (n^2 + |\nabla u|^2)^{-\frac{5}{2}} \times U.$$

Proposition (i) of Theorem 2.3 implies (4.2); hence expression (4.32) is zero. We infer

$$A = 0.$$

Now we let

$$B = (n^2 + |\nabla u|^2)^{-\frac{3}{2}} \times U$$

and claim that

$$B = 0.$$

In fact, if

$$C = \sqrt{2} \cdot |\nabla u| \cdot \sqrt{u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2} + n \cdot |\nabla n|,$$

then inequality (2.3) informs us that C is locally integrable. We have

$$|B| \le \frac{|\nabla u|}{\sqrt{n^2 + |\nabla u|^2}} \times C$$

because of the Cauchy–Schwarz inequality; moreover,

$$A = \frac{|\nabla u|^2}{n^2 + |\nabla u|^2} \times B.$$

Thus A is locally integrable, irrespective of whether it is zero or not; B is locally integrable too, and the following inequality holds:

$$|B| \le |A|^{1/3} \cdot |C|^{2/3},$$

which proves the claim.

Equation (1.5) follows. The proof of Theorem 2.3 is complete. □

5. Remarks on viscosity solutions.

THEOREM 5.1. *A viscosity solution to* (1.5) *is uniquely determined by its boundary values. A smooth solution to* (1.5) *need not be uniquely determined by its boundary values.*

Proof. Theorems 2.2 and 2.3 demonstrate the following property: any viscosity solution to (1.5) which takes the relevant boundary values minimizes the functional J. The former assertion results. The latter results via the analysis of an ad hoc example, as shown below.

Suppose $n \equiv 1$ and u is given by (1.6). (For the sake of brevity, we denote the domain of u by Ω.) Arguments from [26, section 2.2] tell us the following. First, u is a smooth solution to (1.5). Second,

$$-\operatorname{div}\left(\sqrt{1 + |\nabla u|^2}\, \frac{\nabla u}{|\nabla u|}\right)$$

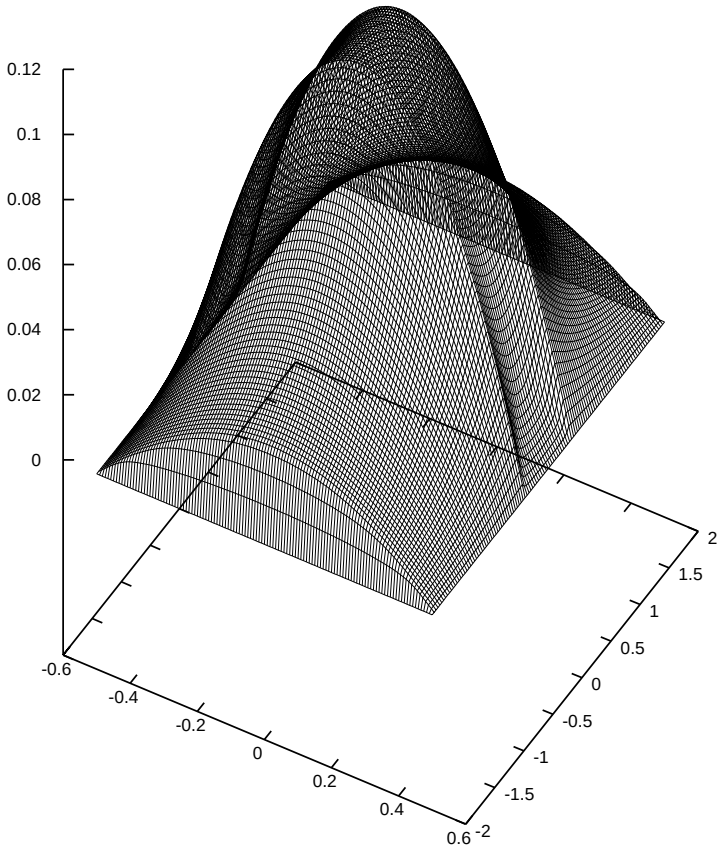


FIG. 5.1. *A plot of the difference between function u given by* (1.6) *and a viscosity solution to* (1.5) *that takes the same boundary values as u.*

equals

$$C_0^\infty(\Omega) \ni \varphi \mapsto 2\int_{-\infty}^{\infty} \varphi(0,y)dy$$

in the sense of distributions. The latter statement informs us that, provided that the domain of J is adjusted as $u + W_0^{1,2}(\Omega)$, the subdifferential of J at u consists of a nonzero measure supported by the y-axis. Therefore, u does not minimize J. It follows that u is not a viscosity solution to (1.5). In other words, u obeys (1.5) but differs from the viscosity solution to (1.5) which is defined in Ω and equals u on $\partial\Omega$. □

Figure 5.1 helps one to visualize the proof above. It displays the difference between the function u given by (1.6) and the viscosity solution to (1.5) that is defined in $]-\frac{1}{2},\frac{1}{2}[\times]-2,2[$ and takes the same values of u on the boundary of such a rectangle. As a matter of fact, such a viscosity solution has been approximated by the solution u_ε to (1.15) with $\varepsilon = 10^{-8}$. Observe the scale in Figure 5.1; we stress that the difference between the u_ε's with $\varepsilon = 10^{-8}$ and $\varepsilon = 10^{-4}$ has order of magnitude 10^{-9}.

Appendix. The following lemma, which analyzes (2.1) closely, is instrumental in proving Theorem 2.2 and proposition (i) of Theorem 2.3.

LEMMA A.1. (i) *f_ε is nonnegative, vanishes only at 0, and is strictly increasing and strictly convex.*

(ii) $$\rho\,\frac{1+\rho^2}{1/2+\rho^2} \le f_\varepsilon'(\rho) \le f'(\rho),$$

and

$$\frac{\rho^2}{2}+\frac{1}{4}\log(1+2\rho^2) \le f_\varepsilon(\rho) \le f(\rho)$$

for every nonnegative ρ.

(iii) *If C_ε is defined by*

$$2C_\varepsilon = \varepsilon^{\frac{1-2\varepsilon}{2(1-\varepsilon)}} + \log(1+\sqrt{\varepsilon}) - \varepsilon - \frac{1}{2}\int_\varepsilon^1 \frac{t^{-\frac{1}{2(1-\varepsilon)}}-\sqrt{t}}{1-t}\,dt,$$

then

$$f_\varepsilon(\rho) = f(\rho) - C_\varepsilon + O(\rho^{-2})$$

as $\rho\to\infty$.

(iv) *f_ε converges uniformly to f on $[0,\infty[$ as ε approaches zero. In effect,*

$$\sup\{|f(\rho)-f_\varepsilon(\rho)| : 0\le\rho<\infty\} = O(\sqrt{\varepsilon}).$$

(v) *f_ε' has a zero of multiplicity one at 0. In effect,*

$$f_\varepsilon'(\rho) = \varepsilon^{-\frac{1}{2(1-\varepsilon)}}\cdot\rho\cdot\left[1-\frac{1}{2\varepsilon}\rho^2+\frac{3+2\varepsilon}{8\varepsilon^2}\rho^4-\frac{15+14\varepsilon+8\varepsilon^2}{48\varepsilon^3}\rho^6+\cdots\right]$$

if $0\le\rho<\sqrt{\varepsilon}$.

(vi) $$\frac{f_\varepsilon'(\rho)}{\rho} < \varepsilon^{-\frac{1}{2(1-\varepsilon)}}$$

and

$$f_\varepsilon''(\rho) < \frac{f_\varepsilon'(\rho)}{\rho}$$

if $\rho>0$;

$$f_\varepsilon''(\rho) \ge \frac{4\varepsilon^{\frac{1-2\varepsilon}{4(1-\varepsilon)}}\,\left[4+\sqrt{\varepsilon(12+\varepsilon)}+\varepsilon\right]}{\left[2+\sqrt{\varepsilon(12+\varepsilon)}+\varepsilon\right]^{\frac{1-2\varepsilon}{2(1-\varepsilon)}}\,\left[\sqrt{12+\varepsilon}+3\sqrt{\varepsilon}\right]^{\frac{3-2\varepsilon}{2(1-\varepsilon)}}}$$

and

$$f_\varepsilon'(\rho)-\rho f_\varepsilon''(\rho) \le \frac{\sqrt{2}\;\varepsilon^{-\frac{\varepsilon}{4(1-\varepsilon)}}\,\left[\sqrt{12+\varepsilon}+\sqrt{\varepsilon}\right]^{\frac{3}{2}}}{\left[2+\sqrt{\varepsilon(12+\varepsilon)}+\varepsilon\right]^{\frac{1-2\varepsilon}{2(1-\varepsilon)}}\left[\sqrt{12+\varepsilon}+3\sqrt{\varepsilon}\right]^{\frac{3-2\varepsilon}{2(1-\varepsilon)}}}$$

if $\rho\ge 0$.

Proof. Equation (2.1) gives successively $f_\varepsilon(0)=0$, and

(A.1) $$f_\varepsilon'(\rho) = \rho\left(\frac{1+\rho^2}{\varepsilon+\rho^2}\right)^{\frac{1}{2(1-\varepsilon)}},$$

$$f_\varepsilon''(\rho) = (1+\rho^2)^{-\frac{1-2\varepsilon}{2(1-\varepsilon)}}(\varepsilon+\rho^2)^{-\frac{3-2\varepsilon}{2(1-\varepsilon)}}(\rho^4+\varepsilon\rho^2+\varepsilon) \tag{A.2}$$

for every nonnegative ρ. Thus $f_\varepsilon'(\rho)$ equals zero if ρ equals zero and is positive if ρ is positive; $f_\varepsilon''(\rho)$ is positive if ρ is nonnegative. Proposition (i) follows.

Equation (A.1) yields

$$\frac{\partial}{\partial\varepsilon}\log f_\varepsilon'(\rho) = \frac{1}{2(1-\varepsilon)^2}\left[\log\left(1+\frac{1-\varepsilon}{\varepsilon+\rho^2}\right) - \frac{1-\varepsilon}{\varepsilon+\rho^2}\right];$$

hence

$$\frac{\partial}{\partial\varepsilon}f_\varepsilon'(\rho) < 0$$

for every positive ρ. In other words, $\varepsilon \mapsto f_\varepsilon'(\rho)$ *decreases* if $\rho > 0$. Since the range of ε is $]0, 1/2]$, proposition (ii) follows. (Incidentally, one might also show that $\varepsilon \mapsto f_\varepsilon'(\rho)$ is *log-convex* for every positive ρ. Observe also that the difference between the r.h.s. and the l.h.s. of the second inequality in (ii) increases as ρ increases from 0 to ∞, approaches $(1+\log 2)/4$ as ρ approaches ∞, and thus is smaller than $0.423286\ldots$.)

Equation (2.1) reads

$$2f_\varepsilon(\rho) = (1-\varepsilon)\int_\varepsilon^{(\varepsilon+\rho^2)/(1+\rho^2)} \frac{t^{\frac{1}{2(1-\varepsilon)}}}{(1-t)^2}\,dt.$$

Integrations by parts and manipulations give

$$2f_\varepsilon(\rho) = (1+\rho^2)^{\frac{1}{2(1-\varepsilon)}}(\varepsilon+\rho^2)^{\frac{1-2\varepsilon}{2(1-\varepsilon)}} - \varepsilon^{\frac{1-2\varepsilon}{2(1-\varepsilon)}}$$

$$+\log\frac{\sqrt{1+\rho^2}+\sqrt{\varepsilon+\rho^2}}{1+\sqrt{\varepsilon}} + \frac{1}{2}\int_\varepsilon^{(\varepsilon+\rho^2)/(1+\rho^2)} \frac{t^{-\frac{1}{2(1-\varepsilon)}} - t^{-\frac{1}{2}}}{1-t}\,dt$$

for every nonnegative ρ. Proposition (iii) follows.

Proposition (ii) ensures that $f - f_\varepsilon$ is nonnegative and increasing, and proposition (iii) ensures that $f(\rho) - f_\varepsilon(\rho)$ approaches C_ε as $\rho \to \infty$. Hence

$$\sup\{|f(\rho) - f_\varepsilon(\rho)| : 0 \le \rho < \infty\} = C_\varepsilon.$$

Proposition (iv) follows.

Proposition (v) follows from manipulations of (A.1).

Equation (A.1) tells us that $f_\varepsilon'(\rho)/\rho$ decreases strictly from $\varepsilon^{-1/(2(1-\varepsilon))}$ to 1 as ρ increases from 0 to ∞. Equations (A.1) and (A.2) imply that

$$\frac{\rho f_\varepsilon''(\rho)}{f_\varepsilon'(\rho)} = 1 - \frac{\rho^2}{(1+\rho^2)(\varepsilon+\rho^2)}$$

if $\rho > 0$ and

$$f_\varepsilon'''(\rho) = (1+\rho^2)^{-\frac{3-4\varepsilon}{2(1-\varepsilon)}}(\varepsilon+\rho^2)^{-\frac{5-4\varepsilon}{2(1-\varepsilon)}}\rho(\rho^4-\varepsilon\rho^2-3\varepsilon)$$

if $\rho \ge 0$. Therefore, $f_\varepsilon''(\rho)$ is less than $f_\varepsilon'(\rho)/\rho$ if ρ is positive; if $\rho = 0$, then $f_\varepsilon'''(\rho)$ and $f_\varepsilon'(\rho) - \rho f_\varepsilon'''(\rho)$ are an absolute maximum and an absolute minimum, respectively; if

$$\rho = \sqrt{\varepsilon/2 + \sqrt{3\varepsilon+\varepsilon^2/4}},$$

then $f_\varepsilon'''(\rho)$ and $f_\varepsilon'(\rho) - \rho f_\varepsilon'''(\rho)$ are an absolute minimum and an absolute maximum, respectively. Proposition (vi) follows. □

Acknowledgment. The authors are indebted to L. Sgheri, a computer scientist of the National Research Council (CNR) of Italy, who developed the relevant algorithms and provided Figures 1.1 and 5.1.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] A. Baernstein, *A unified approach to symmetrization*, in Partial Differential Equations of Elliptic Type, Sympos. Math. 35, A. Alvino, E. Fabes, and G. Talenti, eds., Cambridge University Press, Cambridge, UK, 1994, pp. 47–91.
[3] D. Bouche and F. Molinet, *Méthodes asymptotiques en électromagnétisme*, Springer-Verlag, Berlin, 1994.
[4] S. Choudhary and L. B. Felsen, *Asymptotic theory for inhomogeneous waves*, IEEE Trans. Antennas and Propagation, 21 (1973), pp. 827–842.
[5] S. Choudhary and L. B. Felsen, *Analysis of Gaussian beam propagation and diffraction by inhomogeneous wave tracking*, Proc. IEEE, 62 (1974), pp. 1530–1541.
[6] S. J. Chapman, J. M. H. Lawry, J. R. Ockendon, and R. H. Tew, *On the theory of complex rays*, SIAM Rev., 41 (1999), pp. 417–509.
[7] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
[8] M. G. Crandall, L. C. Evans, and P. L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
[9] B. Dacorogna and P. Marcellini, *Implicit Partial Differential Equations*, Birkhäuser Boston, Boston, 1999.
[10] P. Einzinger and L. B. Felsen, *Evanescent waves and complex rays*, IEEE Trans. Antennas and Propagation, 30 (1982), pp. 594–605.
[11] P. Einzinger and S. Raz, *On the asymptotic theory of inhomogeneous wave tracking*, Radio Science, 15 (1980), pp. 763–771.
[12] L. C. Evans, *On solving certain nonlinear partial differential equations by accretive operator methods*, Israel J. Math., 36 (1980), pp. 225–247.
[13] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.
[14] L. B. Felsen, *Evanescent waves*, J. Opt. Soc. Amer., 66 (1976), pp. 751–760.
[15] L. B. Felsen, *Complex-source-point solutions of the field equations and their relation to the propagation and scattering of Gaussian beams*, in Convegno sulla Teoria Matematica dell'Elettromagnetismo, Sympos. Math. 18, Cambridge University Press, Cambridge, UK, 1976, pp. 39–56.
[16] A. Friedman, *Variational Principles and Free-Boundary Problems*, Wiley, New York, 1982.
[17] M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, Princeton University Press, Princeton, NJ, 1983.
[18] E. Heyman and L. B. Felsen, *Evanescent waves and complex rays for modal propagation in curved open waveguides*, SIAM J. Appl. Math., 43 (1983), pp. 855–884.
[19] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Classics in Appl. Math. 31, SIAM, Philadelphia, 2000.
[20] Yu. A. Kravtsov, *A modification of the geometrical optics method*, Radiofizika, 7 (1964), pp. 664–673.
[21] Yu. A. Kravtsov, *Asymptotic solutions of Maxwell's equations near a caustic*, Radiofizika, 7 (1964), pp. 1049–1056.
[22] O. A. Ladyzenskaja and N. N. Ural'ceva, *Équations aux Dérivées Partielles de Type Elliptique*, Dunod, Paris, 1968.
[23] R. M. Lewis, N. Bleistein, and D. Ludwig, *Uniform asymptotic theory of creeping waves*, Comm. Pure Appl. Math., 20 (1967), pp. 295–328.
[24] D. Ludwig, *Uniform asymptotic expansions at a caustic*, Comm. Pure Appl. Math., 19 (1966), pp. 215–250.
[25] D. Ludwig, *Uniform asymptotic expansion of the field scattered by a convex object at high frequencies*, Comm. Pure Appl. Math., 20 (1967), pp. 103–138.
[26] R. Magnanini and G. Talenti, *On complex-valued solutions to a 2-D eikonal equation. I. Qualitative properties*, in Nonlinear Partial Differential Equations, Contemp. Math. 283, AMS, Providence, RI, 1999, pp. 203–229.
[27] Ch. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, 1966.
[28] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[29] R. T. Rockafellar, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525–539.
[30] C. Rogers, W. K. Schief, and M. E. Johnston, *Bäcklund and his works: Applications in soliton theory*, in Geometric Approaches to Differential Equations, P. J. Vassiliou and I. G. Lisle, eds., Cambridge University Press, Cambridge, UK, 2000, pp. 16–55.
[31] J. T. Schwartz, *Nonlinear Functional Analysis*, Gordon and Breach, New York, 1969.
[32] G. Talenti, *Equazioni lineari ellittiche in due variabili*, Matematiche (Catania), 21 (1966), pp. 339–376.
[33] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* III: *Variational Methods and Optimization*, Springer-Verlag, Berlin, 1985.
[34] W. P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, Berlin, 1989.

SIAM J. MATH. ANAL.
Vol. 34, No. 4, pp. 836–860

NONLINEAR WAVES IN NETWORKS OF NEURONS WITH DELAYED FEEDBACK: PATTERN FORMATION AND CONTINUATION*

LIHONG HUANG† AND JIANHONG WU‡

Abstract. An on-center off-surround network of three identical neurons with delayed feedback is considered, and the effect of synaptic delay of signal transmission on the pattern formation and global continuation of nonlinear waves is described. The spontaneous bifurcation of multiple branches of periodic solutions is discussed, and their spatio-temporal patterns and mode interactions are studied by using the symmetric bifurcation theory of delay differential equations coupled with representation theory of standard dihedral groups, Liapunov's direct method, LaSalle's invariance principle, a priori estimates, and differential inequalities.

Key words. wave, neural network, delay, bifurcation, global continuation

AMS subject classifications. 34K15, 34K20, 34C25, 92B20

PII. S0036141001386519

1. Introduction.

We consider the system of delay differential equations

$$\epsilon\dot{x}_j = -x_j(t) + h(x_j(t-1)) - [g(x_{j-1}(t-1)) + g(x_{j+1}(t-1)) - 2g(x_j(t-1))], \tag{1.1}$$

where $j = 1, 2, 3 (\mathrm{mod}\ 3), \epsilon = \tau^{-1} > 0, \dot{x}_j(t) = \frac{d}{dt}x_j(t), h, g \in C^2(\mathbb{R};\mathbb{R})$ with $h(0) = g(0) = 0$, or, equivalently, we consider

$$\dot{x}_j(t) = -x_j(t) + f(x_j(t-\tau)) - [g(x_{j-1}(t-\tau)) + g(x_{j+1}(t-\tau))] \tag{1.2}$$

with $f = h + 2g$ and $\tau = \epsilon^{-1} > 0$.

Such a system models the evolution of a network of three identical neurons with delayed feedback. There are several reasons why we are particularly interested in such a system. First, if h and g are monotonically increasing, then the network modeled by (1.2) has the property that the self-feedback is excitatory (positive) and the feedback from other neurons is inhibitory (negative). This property is called the on-center off-surround characteristic of a network, and such networks have been found in a variety of neural structures such as neocortex [1], cerebellum [2], and hippocampus [3]. The network described by system (1.2) is of the minimal size among all possible networks with such an on-center off-surround characteristic, and examples of a network of three neurons include the basic rhythm generating circuits of central

*Received by the editors March 16, 2001; accepted for publication (in revised form) July 22, 2002; published electronically February 25, 2003.
http://www.siam.org/journals/sima/34-4/38651.html

†College of Mathematics and Econometrics, Hunan University, Changsha, Hunan 410082, P.R. China (lhhuang@hnu.net.cn). The research of this author was partially supported by the National Natural Science Foundation of P.R. China (10071016), the Doctor Programm Foundation of the Ministry of Education of P.R. China (20010532002), and the Foundation for University Excellent Teacher by the Ministry of Education of P.R. China.

‡Department of Mathematics and Statistics, York University, Toronto, Ontario, M3J 1P3, Canada (wujh@mathstat.yorku.ca). The research of this author was partially supported by the Natural Sciences and Engineering Research Council of Canada, by the Canada Research Chairs Program, and by the Network of Centers of Excellence: Mathematics for Information Technology and Complex Systems.

pattern generators [4, 5] and the canonical cortical circuit proposed in [1, 6]. See also [7] for the motivation of the study of small neural populations. Second, much progress has been made for the theory of dynamics (and, in particular, for the local bifurcation and global continuation of periodic solutions) of scalar delay differential equations (see, for example, [8, 9, 10]), and it is natural to see how the results and methods for scalar delay differential equations can be extended to systems of delay differential equations. Some progress has been made in this direction for a network of two neurons without self-feedback and with delayed interaction (see, for example, [11, 12, 13, 14]). An important factor to the progress in [11, 12, 13] is the fact that such a system can be changed to the so-called unidirectional cyclic system of delay differential equations to which the recently developed powerful theory of Mallet-Paret and Sell [15, 16] and the geometric method developed in [17] can be applied. System (1.2), however, is bidirectional in the sense that the growth rate for the ith cell (component) depends on the feedback from the $(i-1)$th and $(i+1)$th cells, and *both* with a delay. We hope this detailed case study can provide motivation for a more general geometric theory for the global dynamics of bidirectional cyclic systems of delay differential equations. Third, we would like to use this detailed case study to demonstrate how systems with time delay can be used for coupled oscillators. In particular, we note that, in the classical work (see [18] for references), for a ring of cells coupled by diffusion along the sides of a polygon, it was observed that if the coupling is instantaneous, then Hopf bifurcations occur only when the state of each cell is described by at least two variables, and our case study here provides an example in which a ring of cells coupled by delayed nonlinear diffusion exhibits multiple symmetric Hopf bifurcations even when the state of each cell is described by a single variable.

According to the Cohen–Grossberg–Hopfield convergence theorem [19, 20], under standard assumptions on the sigmoid signal functions h and g and if $\tau = 0$, then every solution of system (1.2) is convergent to the set of equilibria. Such a convergence has important applications to a number of areas such as content addressable memory and pattern identification. On the other hand, it was observed in [21] and later confirmed in a number of papers (see [14, 22, 23] and references therein) that large delay may cause nonlinear oscillations in the network. Most of these nonlinear oscillations appear in the form of periodic solutions with certain spatio-temporal structures and, if stable under small perturbation, may represent memory of the network to be stored and retrieved. Therefore, it is important to discuss the spatio-temporal patterns of these periodic solutions and to describe the mode interaction along multiple branches of such periodic solutions.

Needless to say, this is a very difficult task due to the infinite-dimensional nature of the problem caused by the synaptic delay and the possible spatial structure of the system (equivariant with respect to a D_3-action). Some general theorems are available about the existence and global continuation of periodic solutions in symmetric delay differential equations; see [23] for local bifurcation and [24] for global continuation. However, applications of these general results to concrete systems such as (1.1) involve several highly nontrivial tasks: (i) distribution of zeros in characteristic equations which are usually transcendental and depend on parameters; (ii) symmetry analysis on certain generalized eigenspaces of the generator of a linearized system and identification of these spaces with a direct sum of two identical absolute irreducible representations of D_3; (iii) calculation of the so-called crossing numbers which are related to the usual transversality condition in a standard Hopf bifurcation theory (see section 2 for details); (iv) a priori estimation of the period and of the norm of a

periodic solution.

In this paper, we show the following:

(a) The model equation (1.1) is equivariant with respect to a D_3-action.
(b) There exists a sequence of critical values $\{\tau_k\}$ at which the linearization of (1.1) at the zero solution has a pair of purely imaginary eigenvalues.
(c) The generalized eigenspace of the above eigenvalues is four-dimensional and is the direct sum of two identical absolutely irreducible representations of D_3.
(d) Near each τ_k, there exist eight branches of periodic solutions, two of which are phase-locked, three are standing waves, and three are mirror-reflecting waves.
(e) These bifurcations of periodic solutions exist for all $\tau > \tau_k$ (global continuation); the branches of mirror-reflecting waves and the branches of phase-locked oscillations do not coincide, but coincidence of some branches of mirror-reflecting waves and some branches of standing waves may occur through periodic doubling.

The local bifurcation and the asymptotic forms of the aforementioned waves will be described in section 2, and their global continuation will be studied in section 3.

2. The local existence and asymptotic forms of waves. We start by stating a general result due to [23]. Let C denote the Banach space of continuous mappings from $[-1, 0]$ into $\mathbb{R}^n$ equipped with the supremum norm $||\phi|| = \sup_{-1\le\theta\le 0} |\phi(\theta)|$ for $\phi \in C$. In what follows, if $\sigma \in \mathbb{R}$, $A \ge 0$, and $x : [\sigma - 1, \sigma + A] \to \mathbb{R}^n$ is a continuous mapping, then $x_t \in C$, $t \in [\sigma, \sigma + A]$, is defined by $x_t(\theta) = x(t + \theta)$ for $-1 \le \theta \le 0$.

Suppose that $F : C \to \mathbb{R}^n$ is C^2-smooth with $F(0) = 0$. Consider the delay differential equation

$$\dot{x}(t) = \tau F(x_t),$$

where $\tau > 0$. Let $L\phi = DF(0)\phi$ with $\phi \in C$. It is well known that the linear system

$$\dot{x}(t) = \tau L x_t$$

generates a strongly continuous semigroup of linear operators with an infinitesimal generator $A(\tau)$. Moreover, the spectrum $\sigma(A(\tau))$ of $A(\tau)$ consists of eigenvalues which are solutions of the characteristic equation

$$\det \Delta(\tau, \lambda) = 0, \quad \lambda \in \mathbb{C},$$

where $\mathbb{C}$ is the set of all complex numbers, and the characteristic matrix $\Delta(\tau, \lambda)$ is given by

$$\Delta(\tau, \lambda) = \lambda I_n - \tau L(e^{\lambda \cdot} I_n),$$

where I_n is the identity matrix on $\mathbb{C}^n$, $e^{\lambda \cdot} z$ is the mapping from $[-1, 0]$ into $\mathbb{C}^n$ given by $e^{\lambda \cdot} z(\theta) = e^{\lambda\theta} z$ for $z \in \mathbb{C}^n$ and $\theta \in [-1, 0]$, and $L(e^{\lambda \cdot} I_n) = (L(e^{\lambda \cdot} e_1), \dots, L(e^{\lambda \cdot} e_n))$ with $(e_1, \dots, e_n)$ being the standard basis of $\mathbb{R}^n$ and $L(e^{\lambda \cdot} e_j)$ the image of $e^{\lambda \cdot} e_j$ under the complexification of the linear mapping L for each $j = 1, \dots, n$.

We assume the following.

(G1) The characteristic matrix is continuously differentiable in $\tau \in (0, \infty)$, and there exist $\tau_0 \in (0, \infty)$ and $\beta_0 > 0$ such that (i) $A(\tau_0)$ has eigenvalues $\pm i\beta_0$; (ii) the generalized eigenspace, denoted by $U_{(i\beta_0, -i\beta)}(A(\tau_0))$, of these eigenvalues $\pm i\beta_0$

consists of only eigenvectors of $A(\tau_0)$ associated with $\pm i\beta_0$; (iii) all other eigenvalues of $A(\tau_0)$ are not integer multiples of $\pm i\beta_0$.

To state the next assumption that describes the possible (spatial) symmetry of the system considered, we need to introduce some group-theoretic preliminaries. We refer to [18, 25] for more details.

In what follows, by a (compact) Lie group Γ, we mean a closed subgroup of $GL(\mathbb{R}^n)$, the group of all invertible linear transformations of the vector space $\mathbb{R}^n$ into itself. Note that the space of $n\times n$ matrices may be identified with $\mathbb{R}^{n^2}$, which contains $GL(\mathbb{R}^n)$ as an open subset. We say that Γ is a closed subgroup of $GL(\mathbb{R}^n)$ if it is a closed subset of $GL(\mathbb{R}^n)$ as well as a subgroup of $GL(\mathbb{R}^n)$. A specific example of a Lie group is the special orthogonal group $SO(n)$ that consists of all $n \times n$ matrices A such that $AA^T = I_n$ and $\det A = 1$, where A^T is the transpose of A. In particular, $SO(2)$ consists precisely of the planar rotations

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

In this way, $SO(2)$ may be identified with the circle group S^1, the identification being $R_\theta \to e^{i\theta}$. Two other Lie groups will be used in this paper. The first is Z_n, the cyclic group of order n. (The order of a finite group is the number of elements that it contains.) The second is the dihedral group D_n of order $2n$ that is generated by Z_n together with an element (called the flip) of order 2 that does not communicate with Z_n.

Let V be a topological vector space over the field of complex numbers $\mathbb{C}$ or the field of real numbers $\mathbb{R}$, and let $GL(V)$ be the group of isomorphisms of V onto itself. We say that a compact Lie group Γ acts on V if there is a continuous mapping $\Gamma \times V \ni (\gamma, v) \mapsto \gamma \cdot v \in V$ such that (a) for each $\gamma \in \Gamma$, the mapping $\rho_\gamma : V \to V$ given by $\rho_\gamma(v) = \gamma \cdot v$ is linear; (b) if $\gamma_1, \gamma_2 \in \Gamma$, then $\gamma_1 \cdot (\gamma_2 \cdot v) = (\gamma_1\gamma_2) \cdot v$. The mapping that sends $\gamma \in \Gamma$ to $\rho_\gamma \in GL(V)$ is called a representation of Γ on V. In what follows, we shall write γv for $\gamma \cdot v$ for all $\gamma \in \Gamma$ and $v \in V$.

If Γ acts on both V and W and if there is a linear isomorphism $A : V \to W$ such that $A(\gamma v) = \gamma(Av)$ for all $v \in V$ and $\gamma \in \Gamma$, then we say the Γ actions on V and W are isomorphic, and such a linear isomorphism is called a Γ-isomorphism.

Let Γ act on V, and let W be a subspace of V. We say that W is Γ-invariant if $\gamma w \in W$ for every $\gamma \in \Gamma$ and $w \in W$. We thus obtain a Γ-action on W called the restricted action of Γ on W.

Finally, if Γ acts on V, we say a linear mapping $F : V \to V$ is Γ-equivariant if $F(\gamma v) = \gamma F(v)$ for all $\gamma \in \Gamma$ and $v \in V$. A representation of Γ on V is absolutely irreducible if the only linear mappings on V that are Γ-equivariant are scalar multiples of the identity. A Γ-invariant subspace W of V is Γ-irreducible if the only invariant subspaces of W are $\{0\}$ and W. It is known that up to a Γ-isomorphism there are only a finite number of distinct Γ-irreducible subspaces, denoted by $U_1, \dots, U_s$. If we define W_k as the sum of all Γ-irreducible subspaces of V that are Γ-isomorphic to U_k, then $V = W_1 \oplus \cdots W_s$, and this is called an isotypical decomposition of V. In the case in which $\Gamma = Z_N$ and $V = \mathbb{C}^n$ for two fixed positive integers n and N, every irreducible subspace must be one-dimensional, and the restricted action of Γ to any such irreducible subspace is Γ-isomorphic to the Γ-action on $\mathbb{C}$ defined by $\rho \cdot z = \rho^j z$ for some nonnegative integer j and for all $z \in \mathbb{C}$, where ρ is the generator of $Z_N \le S^1$.

With this short introduction to group-theoretic preliminaries, we can now state the next set of assumptions.

(G2) There exists a compact Lie group Γ acting on $\mathbb{R}^n$ such that F is Γ-equivariant; i.e., $F(\gamma\phi) = \gamma F(\phi)$ for $(\gamma, \phi) \in \Gamma \times C$, where $\gamma\phi \in C$ is given by $(\gamma\phi)(\theta) = \gamma\phi(\theta),\ \theta \in [-1, 0]$.

Note that the real Γ-action on $\mathbb{R}^n$ can be naturally extended to a Γ-action on $\mathbb{C}^n$ by

$$\gamma(u + iv) = \gamma u + i\gamma v, \qquad \gamma \in \Gamma, u, v \in \mathbb{R}^n.$$

This action is called the complexification of the Γ-action on $\mathbb{R}^n$. In what follows, we will simply call the complexification of Γ on $\mathbb{C}^n$ the Γ-action on $\mathbb{C}^n$. Due to the Γ-equivariance of F, we can easily show that $\mathrm{Ker}\Delta(\tau_0, i\beta_0)$ is an invariant subspace of $\mathbb{C}^n$ with respect to the complexification of the Γ-action on $\mathbb{R}^n$. We need the following assumption.

(G3) There exists a real m-dimensional absolutely irreducible representation of Γ on V such that the restricted action of Γ on $\mathrm{Ker}\Delta(\tau_0, i\beta_0)$ is isomorphic to the action of Γ on $V \oplus V$ defined by $\gamma(v_1, v_2) = (\gamma v_1, \gamma v_2)$ for $\gamma \in \Gamma, v_1, v_2 \in V$.

Let $\{b_{j1} + ib_{j2}\}_{j=1}^m$ be a basis for $\mathrm{Ker}\Delta(\tau_0, i\beta_0)$, and for any $\beta > 0$ define $\sin_\beta,\ \cos_\beta \in \mathrm{C}([-1, 0]; \mathbb{R})$ by

$$\sin_\beta(\theta) = \sin(\beta\theta),\ \cos_\beta(\theta) = \cos(\beta\theta),\ \theta \in [-1, 0].$$

Then the columns of $\Phi_{\tau_0} = (\varepsilon_1, \ldots, \varepsilon_{2m})$ form a basis for $U_{(i\beta_0, -i\beta_0)}(A(\tau_0))$, where

$$\begin{aligned}
&\varepsilon_j = \sin_{\beta_0} b_{\mathrm{j1}} + \cos_{\beta_0} b_{\mathrm{j2}},\\
&\varepsilon_{m+j} = \cos_{\beta_0} b_{\mathrm{j1}} - \sin_{\beta_0} b_{\mathrm{j2}}, \quad 1 \le j \le m.
\end{aligned}$$

It can be shown (see Lemma 2.1 of [23]) that there exist $\delta_0 > 0$ and a continuously differentiable function $\lambda : (\tau_0 - \delta_0, \tau_0 + \delta_0) \to \mathbb{C}$ such that $\lambda(\tau_0) = i\beta_0,\ \lambda(\tau)$ is an eigenvalue of $A(\tau)$, $U_{(\lambda(\tau), \overline{\lambda(\tau)})}(A(\tau))$ consists of eigenvectors of $A(\tau)$ associated with these eigenvalues, and $\dim U_{(\lambda(\tau), \overline{\lambda(\tau)})}(A(\tau)) = \dim U_{(i\beta_0, -i\beta_0)}(A(\tau_0))$.

We will require the following transversality condition.

(G4) $\frac{d}{d\tau}\mathrm{Re}\lambda(\tau)\,|_{\tau=\tau_0} \neq 0$.

Let $\omega = \frac{2\pi}{\beta_0}$. Denote by P_ω the Banach space of all continuous ω-periodic mappings $x : \mathbb{R} \to \mathbb{R}^n$. Then $\Gamma \times S^1$ acts on P_ω by

$$(\gamma, e^{i\theta})x(t) = \gamma x\left(t + \frac{\theta}{2\pi}\omega\right), \quad (\gamma, e^{i\theta}) \in \Gamma \times S^1,\ x \in P_\omega.$$

Denote by SP_ω the subspace of P_ω consisting of all ω-periodic solutions of $\dot{x}(t) = \tau_0 L x_t$. Then, for each subgroup $\Sigma \le \Gamma \times S^1$, the fixed point set

$$Fix(\Sigma, SP_\omega) = \{x \in SP_\omega; \quad (\gamma, \theta)x = x \text{ for all } (\gamma, \theta) \in \Sigma\}$$

is a subspace.

Under assumption (G1), the columns of $U(t) = \Phi_{\tau_0}(0)e^{B(\tau_0)t},\ t \in \mathbb{R}$, form a basis for SP_ω, where

$$B(\tau_0) = \begin{pmatrix} 0 & -\beta_0 I_m \\ \beta_0 I_m & 0 \end{pmatrix}.$$

Also, SP_ω is a $\Gamma \times S^1$-invariant subspace of P_ω (see Lemma 2.3 of [23]). We can now state the general symmetric local Hopf bifurcation theorem (Theorem 2.1 of [23]).

LEMMA 2.1. *Assume that* (G1)–(G4) *are satisfied and* $\dim Fix(\Sigma, SP_\omega) = 2$ *for some* $\Sigma \le \Gamma \times S^1$. *Then, for a chosen basis* $\{\delta_1, \delta_2\}$ *of* $Fix(\Sigma, SP_\omega)$, *there exist constants* $a_0 > 0$, $\tau_0^* > 0$, $\sigma_0 > 0$, C^1*-smooth functions* $\tau^* : \mathbb{R}^2_{a_0} \to \mathbb{R}$, $\omega^* : \mathbb{R}^2_{a_0} \to (0, \infty)$, *and a* C^1*-smooth mapping* $x^* : \mathbb{R}^2_{a_0} \to C(\mathbb{R}; \mathbb{R}^n)$, *where* $\mathbb{R}^2_{a_0} = \{a \in \mathbb{R}^2; |a| < a_0\}$ *and* $C(\mathbb{R}; \mathbb{R}^n)$ *is the Banach space of all continuous mappings from* $\mathbb{R}$ *into* $\mathbb{R}^n$ *equipped with the supremum norm such that, for each* $a \in \mathbb{R}^2_{a_0}$, $x^*(a)$ *is an* $\omega^*(a)$*-periodic solution of* $\dot{x}(t) = \tau F(x_t)$ *with* $\tau = \tau^*(a)$, *and*

$$\begin{aligned} &\gamma x^*(a)(t) = x^*(a)\left(t - \frac{\omega^*(a)}{\omega}\theta\right), \quad (\gamma, \theta) \in \Sigma, \\ &x^*(0) = 0,\ \omega^*(0) = \omega,\ \tau^*(0) = \tau_0^*, \\ &x^*(a) = (\delta_1, \delta_2)a + o(|a|) \text{ as } |a| \to 0. \end{aligned}$$

Furthermore, for $|\tau - \tau_0| < \tau_0^*$, $|\tilde{\omega} - \frac{2\pi}{\beta_0}| < \sigma_0$, *every* $\tilde{\omega}$*-periodic solution of* $\dot{x}(t) = \tau F(x_t)$ *with* $\|x_t\| < \sigma_0$, $\gamma x(t) = x(t - \frac{\tilde{\omega}}{\omega}\theta)$ *for* $(\gamma, \theta) \in \Sigma$, *and* $t \in \mathbb{R}$ *must be of the above type.*

We now consider the system (1.1). It arises from

$$\dot{y}_j(t) = -y_j(t) + h(y_j(t-\tau)) - [g(y_{j-1}(t-\tau)) + g(y_{j+1}(t-\tau)) - 2g(y_j(t-\tau))]$$

with $\epsilon = \tau^{-1}$ and by the change of variable $x_j(t) = y_j(\tau t)$. We will apply Lemma 2.1 to (1.1) with $F : C \to \mathbb{R}^3$ by

$$(F(\phi))_j = -\phi_j(0) + h(\phi_j(-1)) - [g(\phi_{j-1}(-1)) + g(\phi_{j+1}(-1)) - 2g(\phi_j(-1))]$$

for $\phi \in C := C([-\tau, 0]; \mathbb{R}^3)$ and $j (\mathrm{mod}\ 3)$.

PROPOSITION 2.2. *Let* $\Gamma = D_3$ *be the dihedral group of order* 2×3. *Denote by* ρ *the generator of the cyclic subgroup* $Z_3 \le D_3$ *and by* κ *the flip. Define the action of* Γ *on* $\mathbb{R}^3$ *by*

$$\left\{ \begin{array}{l} (\rho x)_j = x_{j+1}, \quad j (\mathrm{mod}\, 3), \\ (\kappa x)_2 = x_3, (\kappa x)_3 = x_2, (\kappa x)_1 = x_1, x \in \mathbb{R}^3. \end{array} \right. \tag{2.1}$$

Then F *is* Γ*-equivariant.*

Proof. For $\phi \in C$ and $j (\mathrm{mod}\, 3)$, we have

$$\begin{aligned} &(F(\rho\phi))_j \\ &= -(\rho\phi)_j(0) + h((\rho\phi)_j(-1)) - [g((\rho\phi)_{j-1}(-1)) + g((\rho\phi)_{j+1}(-1)) - 2g((\rho\phi)_j(-1))] \\ &= -\phi_{j+1}(0) + h(\phi_{j+1}(-1)) - [g(\phi_j(-1)) + g(\phi_{j+2}(-1)) - 2g(\phi_{j+1}(-1))] \\ &= ((\rho F)(\phi))_j \end{aligned}$$

and

$$\begin{aligned} &(F(\kappa\phi))_1 \\ &= -(\kappa\phi)_1(0) + h((\kappa\phi)_1(-1)) - [g((\kappa\phi)_3(-1)) + g((\kappa\phi)_2(-1)) - 2g((\kappa\phi)_1(-1))] \\ &= -\phi_1(0) + h(\phi_1(-1)) - [g(\phi_2(-1)) + g(\phi_3(-1)) - 2g(\phi_1(-1))] \\ &= ((\kappa F)(\phi))_1. \end{aligned}$$

Moreover,

$$\begin{aligned} &(F(\kappa\phi))_2 \\ &= -(\kappa\phi)_2(0) + h((\kappa\phi)_2(-1)) - [g((\kappa\phi)_1(-1)) + g((\kappa\phi)_3(-1)) - 2g((\kappa\phi)_2(-1))] \\ &= -\phi_3(0) + h(\phi_3(-1)) - [g(\phi_1(-1)) + g(\phi_2(-1)) - 2g(\phi_3(-1))] \\ &= ((\kappa F)(\phi))_2. \end{aligned}$$

Similarly, $(F(\kappa\phi))_3 = ((\kappa F)(\phi))_3$. This completes the proof. □

Let

$$\gamma = h'(0), \qquad \beta = g'(0). \tag{2.2}$$

Then the linearization of (1.1) at $x = 0 \in \mathbb{R}^3$ is

$$\frac{1}{\tau}\dot{X}_j(t) = -X_j(t) + \gamma X_j(t-1) - \beta[X_{j-1}(t-1) + X_{j+1}(t-1) - 2X_j(t-1)], \tag{2.3}$$

where $j = 1, 2, 3 (\mathrm{mod}\, 3)$. The characteristic equation takes the form

$$\det \Delta(\tau, \lambda) = 0,$$

where

$$\Delta(\tau, \lambda) = (\lambda + \tau)I_3 - \tau M e^{-\lambda}, \quad \lambda \in \mathbb{C}, \tag{2.4}$$

and

$$M = \begin{pmatrix} \gamma + 2\beta & -\beta & -\beta \\ -\beta & \gamma + 2\beta & -\beta \\ -\beta & -\beta & \gamma + 2\beta \end{pmatrix}. \tag{2.5}$$

PROPOSITION 2.3. $\det \Delta(\tau, \lambda) = (\lambda + \tau - \gamma\tau e^{-\lambda})[\lambda + \tau - (\gamma + 3\beta)\tau e^{-\lambda}]^2$.

Proof. Let $\chi = e^{i\frac{2\pi}{3}}$ and

$$v_k = (1, \chi^k, \chi^{2k})^T, \quad k = 0, 1, 2. \tag{2.6}$$

Clearly, $v_0 = (1, 1, 1)^T$ and $v_2 = \overline{v}_1$. Let

$$\mathbb{C}_k = \{v_k z; z \in \mathbb{C}\}, \quad k = 0, 1, 2.$$

Then

$$\mathbb{C}^3 = \mathbb{C}_0 \oplus \mathbb{C}_1 \oplus \mathbb{C}_2$$

and

$$\begin{aligned}
&(\Delta(\tau, \lambda)v_k)_j \\
&= (\lambda + \tau - (\gamma + 2\beta)\tau e^{-\lambda})(v_k)_j + \tau\beta e^{-\lambda}(e^{i\frac{2\pi}{3}k} + e^{-i\frac{2\pi}{3}k})(v_k)_j \\
&= \left[\lambda + \tau - \tau(\gamma + 2\beta)e^{-\lambda} + 2\beta\tau\cos\left(\frac{2\pi}{3}k\right)e^{-\lambda}\right](v_k)_j \\
&= \left[\lambda + \tau - \left(\gamma + 4\beta\sin^2\left(\frac{\pi}{3}k\right)\right)\tau e^{-\lambda}\right](v_k)_j.
\end{aligned}$$

That is,

$$\begin{aligned}
&\Delta(\tau, \lambda)|_{\mathbb{C}_k} \\
&= \lambda + \tau - \left(\gamma + 4\beta\sin^2\left(\frac{\pi}{3}k\right)\right)\tau e^{-\lambda} \\
&= \begin{cases} \lambda + \tau - \gamma\tau e^{-\lambda} & \text{if } k = 0, \\ \lambda + \tau - (\gamma + 3\beta)\tau e^{-\lambda} & \text{if } k = 1, 2. \end{cases}
\end{aligned}$$

This completes the proof. □

We now make the following assumption.

(H1) $|\gamma| < 1, \gamma + 3\beta > 1$.

The critical values of τ where the characteristic equation has purely imaginary zeros are described in the following.

PROPOSITION 2.4. *Let $A(\tau)$ denote the infinitesimal generator of the semigroup generated by system* (2.3). *Assume that* (H1) *is satisfied. Define*

$$\begin{cases} \beta_k = 2k\pi - \arccos\ \dfrac{1}{\gamma + 3\beta}, \\ \tau_k = -\beta_k \cot\ \beta_k, \quad k \geq 1. \end{cases}$$

Then the following hold.

(i) *For every fixed $\tau \geq 0$, all zeros of $\lambda + \tau - \gamma\tau e^{-\lambda}$ have negative real parts.*

(ii) *At (and only at) $\tau = \tau_k$, $A(\tau)$ has purely imaginary eigenvalues. These eigenvalues are given by $\pm i\beta_k$ with $\beta_k \in (2k\pi - \frac{\pi}{2}, 2k\pi)$.*

(iii) *All other eigenvalues of $A(\tau_k)$ are not integer multiples of $\pm i\beta_k$.*

(iv) *The generalized eigenspace $U_{(i\beta_k, -i\beta_k)}(A(\tau_k))$ consists of eigenvectors of $A(\tau_k)$ associated with $\pm i\beta_k$ only and*

$$U_{(i\beta_k, -i\beta_k)}(A(\tau_k)) = \left\{ \sum_{i=1}^{4} x_i \epsilon_i; \quad x_i \in \mathbb{R}, i = 1, \dots, 4 \right\},$$

where, for $\theta \in [-1, 0]$,

$$\begin{aligned} \epsilon_1(\theta) &= \mathrm{Re}(e^{i\beta_k\theta} v_1) = \cos(\beta_k\theta)\mathrm{Re}v_1 - \sin(\beta_k\theta)\mathrm{Im}v_1, \\ \epsilon_2(\theta) &= \mathrm{Im}(e^{i\beta_k\theta} v_1) = \sin(\beta_k\theta)\mathrm{Re}v_1 + \cos(\beta_k\theta)\mathrm{Im}v_1, \\ \epsilon_3(\theta) &= \mathrm{Re}(e^{i\beta_k\theta} v_2) = \cos(\beta_k\theta)\mathrm{Re}v_1 + \sin(\beta_k\theta)\mathrm{Im}v_1, \\ \epsilon_4(\theta) &= \mathrm{Re}(e^{i\beta_k\theta} v_2) = \sin(\beta_k\theta)\mathrm{Re}v_1 - \cos(\beta_k\theta)\mathrm{Im}v_1. \end{aligned}$$

Proof. (i) Let $\lambda = u + iv$ be a zero of $\lambda + \tau - \gamma\tau e^{-\lambda}$. Then we get $v = -\gamma\tau e^{-u}\mathrm{sin}v$ and $u + \tau = \gamma\tau e^{-u}\mathrm{cos}v$, from which it follows that

$$\gamma^2\tau^2 e^{-2u} = v^2 + (u + \tau)^2.$$

Consequently, $u < 0$, for otherwise the left-hand side of the above equality is strictly less than τ^2, while the right-hand side is larger than or equal to τ^2.

To verify (ii)–(iv), let $\lambda = iv$ with $v > 0$ be a solution of $\lambda + \tau - (\gamma + 3\beta)\tau e^{-\lambda} = 0$. Then

$$\begin{cases} \tau = (\gamma + 3\beta)\tau \mathrm{cos}v, \\ v = -(\gamma + 3\beta)\tau \mathrm{sin}v. \end{cases}$$

So

$$\tan v = -\frac{v}{\tau},$$

from which it follows that $\tan v < 0$, and hence $v \notin [0, \frac{\pi}{2}] + \mathbb{Z}\pi$; here $\mathbb{Z}$ is the set of all integers. Therefore, we must have

$$v = 2k\pi - \arccos \frac{1}{\gamma + 3\beta} = \beta_k, \quad k \geq 1,$$

and

$$\tau = -\beta_k \cot \beta_k = \tau_k.$$

Therefore, $\lambda + \tau - (\gamma + 3\beta)\tau e^{-\lambda} = 0$ has purely imaginary roots (given by $i\beta_k$) if and only if $\tau = \tau_k$ for some $k \geq 1$.

It is well known that $\phi \in C([-1, 0]; \mathbb{C}^3)$ is an eigenvector of $A(\tau_k)$ associated with the eigenvalue $i\beta_k$ if and only if $\phi(\theta) = e^{i\beta_k\theta}z, -1 \leq \theta \leq 0$, for some vector $z \in \mathbb{C}^3$ such that $\Delta(\tau_k, i\beta_k)z = 0$ (see, for example, pp. 198 in [9]). From the proof of Proposition 2.3, we then have $v \in \langle v_1, v_2\rangle$, the complex space spanned by v_1 and v_2. Similar arguments apply to $-i\beta_k$. Therefore, the eigenspace of $A(\tau_k)$ associated with $\pm i\beta_k$ is spanned by $e^{i\beta_k\theta}v_1, e^{i\beta_k\theta}v_2, e^{-i\beta_k\theta}v_1$, and $e^{-i\beta_k\theta}v_2$. Therefore, this space has the real basis $\{\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$. On the other hand, the eigenspace of $A(\tau_k)$ associated with $i\beta_k$ is of dimension 2 and the algebraic multiplicity of $\lambda = i\beta_k$ as a zero of $\det \Delta(\tau_k, \lambda) = 0$ is also 2. So the well-known folk theorem in functional differential equations (see [26] or Theorem 4.2 in [9]) implies that $U_{(i\beta_k, -i\beta_k)}(A(\tau_k))$ must coincide with the eigenspace of $A(\tau_k)$ associated with $\pm i\beta_k$. This completes the proof. □

PROPOSITION 2.5. *Let $\Gamma = D_3$ act on $\mathbb{R}^2$ by*

$$\rho\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$
$$\kappa\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ -x_2 \end{pmatrix}, \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2.$$

Then $\mathbb{R}^2$ is an absolutely irreducible representation of Γ, and the restricted action of Γ on $\mathrm{Ker}\Delta(\tau_k, i\beta_k)$ is isomorphic to the action of Γ on $\mathbb{R}^2 \oplus \mathbb{R}^2$.

Proof. The proof for the absolute irreducibility of the representation of Γ on $\mathbb{R}^2$ is straightforward and can be found in, for example, [18]. Clearly,

$$\mathrm{Ker}\Delta(\tau_k, i\beta_k) = \{(x_1 + ix_2)v_1 + (x_3 + ix_4)v_2; \quad x_i \in \mathbb{R}, i = 1, \ldots, 4\}.$$

Define

$$J((x_1 + ix_2)v_1 + (x_3 + ix_4)v_2) = (x_1 + x_3, x_2 - x_4, x_2 + x_4, x_3 - x_1)^T.$$

Clearly, $J : \mathrm{Ker}\Delta(\tau_k, i\beta_k) \cong \mathbb{R}^4$ is a linear isomorphism. Note that

$$\begin{aligned}
&\rho[(x_1 + ix_2)v_1 + (x_3 + ix_4)v_2] \\
&= (x_1 + ix_2)e^{i\frac{2\pi}{3}}v_1 + (x_3 + ix_4)e^{-i\frac{2\pi}{3}}v_2 \\
&= \left[\left(-\frac{1}{2}x_1 - \frac{\sqrt{3}}{2}x_2\right) + i\left(-\frac{1}{2}x_2 + \frac{\sqrt{3}}{2}x_1\right)\right]v_1 \\
&\quad + \left[\left(-\frac{1}{2}x_3 + \frac{\sqrt{3}}{2}x_4\right) + i\left(-\frac{1}{2}x_4 - \frac{\sqrt{3}}{2}x_3\right)\right]v_2
\end{aligned}$$

and

$$\kappa[(x_1 + ix_2)v_1 + (x_3 + ix_4)v_2] = (x_1 + ix_2)v_2 + (x_3 + ix_4)v_1.$$

Therefore,

$$\begin{aligned}
&J(\rho[(x_1+ix_2)v_1+(x_3+ix_4)v_2])\\
&=\left(-\frac{1}{2}(x_1+x_3)-\frac{\sqrt{3}}{2}(x_2-x_4),-\frac{1}{2}(x_2-x_4)+\frac{\sqrt{3}}{2}(x_1+x_3),\right.\\
&\qquad\left.-\frac{1}{2}(x_2+x_4)-\frac{\sqrt{3}}{2}(x_3-x_1),-\frac{1}{2}(x_3-x_1)+\frac{\sqrt{3}}{2}(x_2+x_4)\right)^T\\
&=\rho J((x_1+ix_2)v_1+(x_3+ix_4)v_2)
\end{aligned}$$

and

$$\begin{aligned}
&J(\kappa[(x_1+ix_2)v_1+(x_3+ix_4)v_2])\\
&=(x_3+x_1,x_4-x_2,x_4+x_2,x_1-x_3)^T\\
&=\kappa J[(x_1+ix_2)v_1+(x_3+ix_4)v_2].
\end{aligned}$$

This completes the proof. ∎

PROPOSITION 2.6. *For each fixed $k\geq 1$, there exist $\delta_k>0$ and a C^1-mapping $\lambda_k:(\tau_k-\delta_k,\tau_k+\delta_k)\to\mathbb{C}$ such that $\lambda_k(\tau_k)=i\beta_k$ and $\lambda_k(\tau)+\tau-(\gamma+3\beta)\tau e^{-\lambda_k(\tau)}=0$ for all $\tau\in(\tau_k-\delta_k,\tau_k+\delta_k)$. Moreover, $\frac{d}{d\tau}\mathrm{Re}\lambda_k(\tau)|_{\tau=\tau_k}>0$.*

Proof. The existence of δ_k and the mapping λ_k follow from the implicit function theorem. We now substitute $\lambda=\lambda_k(\tau)$ into $\lambda+\tau-(\gamma+3\beta)\tau e^{-\lambda}=0$, differentiating the equality with respect to τ, to get

$$\begin{aligned}
&\frac{d}{d\tau}\mathrm{Re}\lambda_k(\tau)|_{\tau=\tau_k}\\
&=\mathrm{Re}\frac{-1+(\gamma+3\beta)e^{-\lambda}}{1+\tau(\gamma+3\beta)e^{-\lambda}}\bigg|_{\lambda=i\beta_k,\tau=\tau_k}\\
&=\mathrm{Re}\frac{\lambda/\tau}{1+(\lambda+\tau)}\bigg|_{\lambda=i\beta_k,\tau=\tau_k}\\
&=\frac{\beta_k^2}{\tau_k[(1+\tau_k)^2+\beta_k^2]}.
\end{aligned}$$

This completes the proof. ∎

Fix $k\geq 1$. Let $\omega=\frac{2\pi}{\beta_k}$, and let P_ω be the Banach space of continuous ω-periodic mappings $x:\mathbb{R}\to\mathbb{R}^3$. $\Gamma\times S^1$ acts on P_ω by

$$(\gamma,e^{i\theta})x(t)=\gamma x(t+\theta),\quad e^{i\theta}\in S^1, x\in P_\omega,\gamma\in\Gamma.$$

We will write γx for $(\gamma,1)x$ when $\gamma\in\Gamma$ and $x\in P_\omega$. Let SP_ω denote the subspace of P_ω consisting of all ω-periodic solutions of (2.3) with $\tau=\tau_k$. Then

$$SP_\omega=\{x_1\epsilon_1^*+x_2\epsilon_2^*+x_3\epsilon_3^*+x_4\epsilon_4^*;\quad x_i\in\mathbb{R}, i=1,\dots,4\},$$

where

$$\begin{cases}
\epsilon_1^*(t)=\cos(\beta_k t)w_1-\sin(\beta_k t)w_2,\\
\epsilon_2^*(t)=\sin(\beta_k t)w_1+\cos(\beta_k t)w_2,\\
\epsilon_3^*(t)=\cos(\beta_k t)w_1+\sin(\beta_k t)w_2,\\
\epsilon_4^*(t)=\sin(\beta_k t)w_1-\cos(\beta_k t)w_2,
\end{cases}$$

and

$$w_1 = \left(1, -\frac{1}{2}, -\frac{1}{2}\right)^T, \quad w_2 = \left(0, \frac{\sqrt{3}}{2}, -\frac{\sqrt{3}}{2}\right)^T.$$

PROPOSITION 2.7. *With ϵ_i^* given above, we have*

(i) $\kappa\epsilon_1^* = \epsilon_3^*, \kappa\epsilon_2^* = \epsilon_4^*, \kappa\epsilon_3^* = \epsilon_1^*, \kappa\epsilon_4^* = \epsilon_2^*$;

(ii) $\rho\epsilon_1^* = -\frac{1}{2}\epsilon_1^* - \frac{\sqrt{3}}{2}\epsilon_2^*, \rho\epsilon_2^* = -\frac{1}{2}\epsilon_2^* + \frac{\sqrt{3}}{2}\epsilon_1^*, \rho\epsilon_3^* = -\frac{1}{2}\epsilon_3^* + \frac{\sqrt{3}}{2}\epsilon_4^*, \rho\epsilon_4^* = -\frac{1}{2}\epsilon_4^* - \frac{\sqrt{3}}{2}\epsilon_3^*$.

Proof. (i) is obvious from the definition of the action of κ in Proposition 2.2. To prove (ii), we note that

$$\rho\begin{pmatrix} 1 \\ -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ 1 \end{pmatrix} = -\frac{1}{2}\begin{pmatrix} 1 \\ -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix} - \frac{\sqrt{3}}{2}\begin{pmatrix} 0 \\ \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} \end{pmatrix},$$

$$\rho\begin{pmatrix} 0 \\ \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} \\ 0 \end{pmatrix} = \frac{\sqrt{3}}{2}\begin{pmatrix} 1 \\ -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 0 \\ \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} \end{pmatrix}.$$

So

$$\rho\epsilon_1^* = \cos(\beta_k t)\left[-\frac{1}{2}w_1 - \frac{\sqrt{3}}{2}w_2\right] - \sin(\beta_k t)\left[\frac{\sqrt{3}}{2}w_1 - \frac{1}{2}w_2\right] = -\frac{1}{2}\epsilon_1^* - \frac{\sqrt{3}}{2}\epsilon_2^*,$$

$$\rho\epsilon_2^* = \sin(\beta_k t)\left[-\frac{1}{2}w_1 - \frac{\sqrt{3}}{2}w_2\right] + \cos(\beta_k t)\left[\frac{\sqrt{3}}{2}w_1 - \frac{1}{2}w_2\right] = \frac{\sqrt{3}}{2}\epsilon_1^* - \frac{1}{2}\epsilon_2^*,$$

$$\rho\epsilon_3^* = \cos(\beta_k t)\left[-\frac{1}{2}w_1 - \frac{\sqrt{3}}{2}w_2\right] + \sin(\beta_k t)\left[\frac{\sqrt{3}}{2}w_1 - \frac{1}{2}w_2\right] = -\frac{1}{2}\epsilon_3^* + \frac{\sqrt{3}}{2}\epsilon_4^*,$$

$$\rho\epsilon_4^* = \sin(\beta_k t)\left[-\frac{1}{2}w_1 - \frac{\sqrt{3}}{2}w_2\right] - \cos(\beta_k t)\left[\frac{\sqrt{3}}{2}w_1 - \frac{1}{2}w_2\right] = -\frac{\sqrt{3}}{2}\epsilon_3^* - \frac{1}{2}\epsilon_4^*.$$

This completes the proof. □

Note that, if x is a periodic solution of (1.1), then so is $(\gamma, e^{i\theta})x$ for every $(\gamma, e^{i\theta}) \in \Gamma \times S^1$. If the symmetry of x is Σ_x for a subgroup of $\Gamma \times S^1$, that is, $\Sigma_x = \{(\gamma, e^{i\theta}) \in \Gamma \times S^1; (\gamma, e^{i\theta})x = x\}$, then the symmetry of $(\gamma, e^{i\theta})x$ is given by $(\gamma, e^{i\theta})\Sigma_x(\gamma, e^{i\theta})^{-1}$, which is conjugate to Σ_x. It is known that the subgroups of $D_3 \times S^1$, up to conjugacy, that describe the symmetry of periodic solutions of (1.1) which exhibit certain spatial-temporal patterns are given below (see, for example, p. 368 in [18]):

$$\Sigma_{(2,3)}^{\pm} = \langle(\kappa, \pm 1)\rangle,$$
$$\Sigma_{\rho}^{\pm} = \langle(\rho, e^{\pm i\frac{2\pi}{3}})\rangle.$$

More specifically, for example, $\Sigma_{(2,3)}^{-}$ is a group generated by $(\kappa, -1) \in D_3 \times S^1$.

PROPOSITION 2.8.

$$Fix(\Sigma_{(2,3)}^{+}, SP_\omega) = \{y\cos(\beta_k t)w_1 + z\sin(\beta_k t)w_1; \quad y, z \in \mathbb{R}\},$$
$$Fix(\Sigma_{(2,3)}^{-}, SP_\omega) = \{y\cos(\beta_k t)w_2 + z\sin(\beta_k t)w_2; \quad y, z \in \mathbb{R}\},$$
$$Fix(\Sigma_{\rho}^{-}, SP_\omega) = \{y\epsilon_1^* + z\epsilon_2^*; \quad y, z \in \mathbb{R}\},$$
$$Fix(\Sigma_{\rho}^{=}, SP_\omega) = \{y\epsilon_3^* + z\epsilon_4^*; \quad y, z \in \mathbb{R}\}.$$

Proof. First, $x \in Fix(\Sigma^+_{(2,3)}, SP_\omega)$ if and only if $\kappa x = x$. However, for $x = \sum_{i=1}^4 x_i\epsilon_i^*$, we have

$$\kappa x = x_1\epsilon_3^* + x_2\epsilon_4^* + x_3\epsilon_1^* + x_4\epsilon_2^*.$$

Therefore, $x \in Fix(\Sigma^+_{(2,3)}, SP_\omega)$ if and only if $x_1 = x_3$ and $x_2 = x_4$. This shows that $Fix(\Sigma^+_{(2,3)}, SP_\omega)$ is spanned by $\epsilon_1^* + \epsilon_3^*$ and $\epsilon_2^* + \epsilon_4^*$.

Second, $x \in Fix(\Sigma^-_{(2,3)}, SP_\omega)$ if and only if $\kappa x(t) = x(t + \frac{\omega}{2})$ for $t \in \mathbb{R}$. Let $x = \sum_{i=1}^4 x_i\epsilon_i^*$. Then, as $\cos(\beta_k t + \beta_k \frac{\omega}{2}) = -\cos(\beta_k t)$ and $\sin(\beta_k t + \beta_k \frac{\omega}{2}) = -\sin(\beta_k t)$, we get $\epsilon_i^*(t + \frac{\omega}{2}) = -\epsilon_i^*(t)$, and thus $x(t + \frac{\omega}{2}) = -\sum_{i=1}^4 x_i\epsilon_i^*$. This implies that $\kappa x(t) = x(t + \frac{\omega}{2})$ if and only if $x_1 = -x_3$ and $x_2 = -x_4$. Therefore, $Fix(\Sigma^-_{(2,3)}, SP_\omega)$ is spanned by $\epsilon_1^* - \epsilon_3^*$ and $\epsilon_2^* - \epsilon_4^*$.

Third, for $x = \sum_{i=1}^4 x_i\epsilon_i^*$, we have

$$\begin{aligned}\rho x = &\left(-\frac{1}{2}x_1 + \frac{\sqrt{3}}{2}x_2\right)\epsilon_1^* + \left(-\frac{\sqrt{3}}{2}x_1 - \frac{1}{2}x_2\right)\epsilon_2^* \\ &+ \left(-\frac{1}{2}x_3 - \frac{\sqrt{3}}{2}x_4\right)\epsilon_3^* + \left(\frac{\sqrt{3}}{2}x_3 - \frac{1}{2}x_4\right)\epsilon_4^*.\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\cos\left(\beta_k\left(t \pm \frac{\omega}{3}\right)\right) &= \cos\left(\beta_k t \pm \frac{2\pi}{3}\right) = -\frac{1}{2}\cos(\beta_k t) \mp \frac{\sqrt{3}}{2}\sin(\beta_k t), \\ \sin\left(\beta_k\left(t \pm \frac{\omega}{3}\right)\right) &= \sin\left(\beta_k t \pm \frac{2\pi}{3}\right) = \pm\frac{\sqrt{3}}{2}\cos(\beta_k t) - \frac{1}{2}\sin(\beta_k t).\end{aligned}$$

This, together with the expression of each ϵ_i^* and $x = \sum_{i=1}^4 x_i\epsilon_i^*$, leads to

$$\begin{aligned}&x\left(t \pm \frac{\omega}{3}\right) \\ &= \left(-\frac{1}{2}x_1 \pm \frac{\sqrt{3}}{2}x_2\right)\epsilon_1^* + \left(\mp\frac{\sqrt{3}}{2}x_1 - \frac{1}{2}x_2\right)\epsilon_2^* \\ &\quad + \left(-\frac{1}{2}x_3 \pm \frac{\sqrt{3}}{2}x_4\right)\epsilon_3^* + \left(\mp\frac{\sqrt{3}}{2}x_3 - \frac{1}{2}x_4\right)\epsilon_4^*.\end{aligned}$$

Thus $x \in Fix(\Sigma^\pm_\rho, SP_\omega)$, i.e., $\rho x(t \pm \frac{\omega}{3}) = x(t)$, if and only if

$$\begin{cases} -\frac{1}{2}x_1 + \frac{\sqrt{3}}{2}x_2 = -\frac{1}{2}x_1 \mp \frac{\sqrt{3}}{2}x_2, \\ -\frac{\sqrt{3}}{2}x_1 - \frac{1}{2}x_2 = \pm\frac{\sqrt{3}}{2}x_1 - \frac{1}{2}x_2, \\ -\frac{1}{2}x_3 - \frac{\sqrt{3}}{2}x_4 = -\frac{1}{2}x_3 \mp \frac{\sqrt{3}}{2}x_4, \\ \frac{\sqrt{3}}{2}x_3 - \frac{1}{2}x_4 = \pm\frac{\sqrt{3}}{2}x_3 - \frac{1}{2}x_4. \end{cases}$$

That is, $\rho x(t) = x(t+\frac{\omega}{3})$ if and only if $x_3 = x_4 = 0$, and $\rho x(t) = x(t-\frac{\omega}{3})$ if and only if $x_1 = x_2 = 0$. Therefore, $Fix(\Sigma_\rho^-, SP_\omega)$ is spanned by ϵ_1^* and ϵ_2^*, and $Fix(\Sigma_\rho^+, SP_\omega)$ is spanned by ϵ_3^* and ϵ_4^*. This completes the proof. □

We can now apply Lemma 2.1 to obtain the following main result of this section.

THEOREM 2.9. *Assume that* (H1) *is satisfied. Then, near $\tau = \tau_k$ for each $k \geq 1$, system* (1.1) *has eight distinct branches of periodic solutions bifurcated from the trivial solution $x = 0$. More precisely, we have the following.*

(i) *There exist $\epsilon_0^m > 0$ and $\delta_0^m > 0$ such that, for each $\theta \in [0, 2\pi], \alpha \in (0, \epsilon_0^m)$, system* (1.1) *with $\tau = \tau_k + \tau^m(\alpha, \theta)$ has a periodic solution $x^m = x^m(t; \alpha, \theta)$ with period $\omega^m(\alpha, \theta)$ such that*

$$x_2^m(t;\alpha,\theta) = x_3^m(t;\alpha,\theta),$$
$$x^m(t;\alpha,\theta) = \alpha\cos(\beta_k t + \theta)\left(1, -\frac{1}{2}, -\frac{1}{2}\right)^{\mathrm{T}} + o(|\alpha|) \text{ as } \alpha \to 0.$$

The mapping $(x^m, \tau^m, \omega^m) : (0, \epsilon^m) \times [0, 2\pi] \to C(\mathbb{R}; \mathbb{R}^3) \times \mathbb{R} \times \mathbb{R}$ is C^1-smooth, and

$$\omega^m(0,\theta) = \frac{2\pi}{\beta_k}, \quad \tau^m(0,\theta) = 0.$$

Furthermore, if $|\tau - \tau_k| < \delta_0^m$ and $|\omega - \frac{2\pi}{\beta_k}| < \delta_0^m$, then every ω-periodic solution of (1.1) *satisfying $x_2(t) = x_3(t)$ and $\sup_{t\in\mathbb{R}} |x(t)| < \delta_0^m$ must be given by $x^m(t; \alpha, \theta)$ for some $\alpha \in (0, \epsilon_0^m)$ and $\theta \in [0, 2\pi)$. Similar results hold when we replace $(2, 3)$ by $(1, 2)$ or $(1, 3)$.*

(ii) *There exist $\epsilon_0^s > 0$ and $\delta_0^s > 0$ such that, for each $\theta \in [0, 2\pi], \alpha \in (0, \epsilon_0^s)$, system* (1.1) *with $\tau = \tau_k + \tau^s(\alpha, \theta)$ has a periodic solution $x^s = x^s(t; \alpha, \theta)$ with period $\omega^s = \omega^s(\alpha, \theta)$ such that*

$$x_1^s(t) = x_1^s\left(t - \frac{\omega^s}{2}\right), x_2^s(t) = x_3^s\left(t - \frac{\omega^s}{2}\right), x_3^s(t) = x_3^s(t + \omega^s),$$
$$x^s(t;\alpha,\theta) = \alpha\cos(\beta_k t + \theta)\left(0, -\frac{\sqrt{3}}{2}, -\frac{\sqrt{3}}{2}\right)^{\mathrm{T}} + o(|\alpha|) \text{ as } \alpha \to 0.$$

The mapping $(x^s, \tau^s, \omega^s) : (0, \epsilon_0^s) \times [0, 2\pi] \to C(\mathbb{R}; \mathbb{R}^3) \times \mathbb{R} \times \mathbb{R}$ is C^1-smooth, and

$$\omega^s(0,\theta) = \frac{2\pi}{\beta_k}, \quad \tau^s(0,\theta) = 0.$$

Furthermore, if $|\tau - \tau_k| < \delta_0^s$ and $|\omega - \frac{2\pi}{\beta_k}| < \delta_0^s$, then every ω-periodic solution of (1.1) *satisfying $x_1(t) = x_1(t - \frac{\omega}{2}), x_2(t) = x_3(t - \frac{\omega}{2})$, and $\sup_{t\in\mathbb{R}} |x(t)| < \delta_0^s$ must be given by $x^s(t; \alpha, \theta)$ for some $\alpha \in (0, \epsilon_0^s)$ and $\theta \in [0, 2\pi)$. Similar results hold when we replace $(1, 2, 3)$ by $(2, 1, 3)$ or $(3, 2, 1)$.*

(iii) *There exist $\epsilon_0^d > 0$ and $\delta_0^d > 0$ such that, for each $\theta \in [0, 2\pi], \alpha \in (0, \epsilon_0^d)$, system* (1.1) *with $\tau = \tau_k + \tau^d(\alpha, \theta)$ has a periodic solution $x^d = x^d(t; \alpha, \theta)$ with period $\omega^d = \omega^d(\alpha, \theta)$ such that*

$$x_1^d(t) = x_2^d\left(t \pm \frac{\omega^d}{3}\right), x_2^d(t) = x_3^d\left(t \pm \frac{\omega^d}{3}\right),$$
$$x^d(t;\alpha,\theta) = \alpha\left(\cos(\beta_k t + \theta), \cos\left(\beta_k t + \theta \mp \frac{2\pi}{3}\right), \cos\left(\beta_k t + \theta \mp \frac{4\pi}{3}\right)\right)^T + o(|\alpha|)$$

as $\alpha \to 0$. *The mapping* $(x^d, \tau^d, \omega^d) : (0, \epsilon_0^d) \times [0, 2\pi] \to C(\mathbb{R}; \mathbb{R}^3) \times \mathbb{R} \times \mathbb{R}$ *is* C^1*-smooth, and*

$$\omega^d(0, \theta) = \frac{2\pi}{\beta_k}, \quad \tau^d(0, \theta) = 0.$$

Furthermore, if $|\tau - \tau_k| < \delta_0^d$ *and* $|\omega - \frac{2\pi}{\beta_k}| < \delta_0^d$, *then every* ω*-periodic solution of* (1.1) *satisfying* $x_1(t) = x_2(t \pm \frac{\omega}{3}), x_2(t) = x_3(t \pm \frac{\omega}{3})$, *and* $\sup_{t \in \mathbb{R}} |x(t)| < \delta_0^d$ *must be given by* $x^d(t; \alpha, \theta)$ *for some* $\alpha \in (0, \epsilon_0^d)$ *and* $\theta \in [0, 2\pi)$. *Similar results hold when we replace* $(1, 2, 3)$ *by* $(2, 1, 3)$ *or* $(3, 2, 1)$.

We call periodic solutions in (i)–(iii) *mirror-reflecting waves, standing waves*, and *discrete waves*, respectively. Note that Theorem 2.9 does not rule out the case in which $\tau^l(\alpha, \theta) \leq 0$ $(l = m, s, d)$. In next section, we will use the global bifurcation theory to rule out this case. In fact, we will show that all eight branches of waves are *supercritical* and *global*; i.e., all eight branches of waves exist for $\tau > \tau_k$.

3. Global continuation of waves. We will need a general global symmetric Hopf bifurcation theorem developed in [24]. Namely, we consider the one-parameter family of retarded functional differential equations

$$\dot{x}(t) = \tau F(x_t), \tag{3.1}$$

where $x \in \mathbb{R}^n$, $\tau \in (0, \infty)$, and $F : C([-\tau, 0]; \mathbb{R}^n) \to \mathbb{R}^n$ is continuously differentiable and completely continuous. Furthermore, we assume the following.

(A1) $\Gamma := Z_N$ for some integer N acts on $\mathbb{R}^n$ and $F : C \to \mathbb{R}^n$ is Γ-equivariant.

(A2) For every $x_0 \in M^\Gamma := \{x \in \mathbb{R}^n; \gamma x = x \text{ for } \gamma \in \Gamma, F(\bar{x}) = 0\}$, where $\bar{x} \in C$ is the constant mapping with the constant value $x \in \mathbb{R}^n$, $\det D\hat{F}(x_0) \neq 0$, where $\hat{F}$ is the C^1 mapping from $\mathbb{R}^n$ into $\mathbb{R}^n$, induced by F according to $\hat{F}(x) = F(\bar{x})$ for $x \in \mathbb{R}^n$.

(A3) For every $\tau_0 > 0$ and $x_0 \in M^\Gamma$ such that the generator $A(\tau_0, x_0)$ of the linearized system of (3.1) with $\tau = \tau_0$ at $x = x_0$ has a pair of purely imaginary eigenvalues $\pm i\beta_0$, there exist positive constants b, c, and δ such that (i) the only possible eigenvalue $u + iv$ of $A(\tau_0, x_0)$ with $(u, v) \in \partial\Omega$ is $i\beta_0$, where $\Omega := (0, b) \times (\beta_0 - c, \beta_0 + c)$; (ii) for $(\tau, \beta) \in [\tau_0 - \delta, \tau_0 + \delta] \times [\beta_0 - c, \beta_0 + c]$, $i\beta$ is an eigenvalue of $A(\tau, x_0)$ if and only if $\tau = \tau_0, \beta = \beta_0$.

(A4) $M^* := \{(\tau, x, \beta) \in (0, \infty) \times M^\Gamma \times (0, \infty); \pm i\beta \text{ are eigenvalues of } A(\tau, x)\}$ is a discrete set.

Note that the action of Γ on $\mathbb{R}^n$ induces an action on $\mathbb{C}^n = \mathbb{R}^n + i\mathbb{R}^n$, with respect to which we have the isotypical decomposition

$$\mathbb{C}^n = \mathbb{C}_0^n \oplus \mathbb{C}_1^n \oplus \cdots \oplus \mathbb{C}_j^n \oplus \cdots,$$

where $\mathbb{C}_j^n, j \geq 0$, is the direct sum of all one-dimensional Γ-irreducible subspaces V of $\mathbb{C}^n$ such that the restricted action Γ on V is isomorphic to the Γ-action on $\mathbb{C}$ defined by $\rho \cdot z = \rho^j z$ for the generator $\rho \in Z_N \leq S^1$ and for $z \in \mathbb{C}$. Let

$$\Delta_{x_0}(\tau, \lambda) := \lambda I_n - \tau D_\phi F(\bar{x}_0)(e^{\lambda \cdot} I_n) \tag{3.2}$$

for $\tau > 0, x_0 \in M^\Gamma$, and $\lambda \in \mathbb{C}$. By assumption (A1), we have $\Delta_{x_0}(\tau, \lambda)\mathbb{C}_j^n \subset \mathbb{C}_j^n$ for $j \geq 0$ and for $\lambda \in \mathbb{C}$. Put

$$\Delta_{x_0, j}(\tau, \lambda) = \Delta_{x_0}(\tau, \lambda)|_{\mathbb{C}_j^n}, \quad j \geq 0. \tag{3.3}$$

Clearly, $\Delta_{x_0}(\tau,\lambda)$ is analytic in $\lambda\in\mathbb{C}$ and continuous in $\tau>0$. So, under assumption (A3), we may assume that $\det\Delta_{x_0}(\tau_0\pm\delta,u+iv)\neq 0$ for $(u,v)\in\partial\Omega$. Therefore, $\det\Delta_{x_0,j}(\tau_0\pm\delta,u+iv)\neq 0$ for $(u,v)\in\partial\Omega$ and for $j\geq 0$. Consequently, the following integers are well defined:

$$c_j(x_0,\tau_0,\beta_0)=\deg_B(\det\Delta_{x_0,j}(\tau_0-\delta,\cdot),\Omega)-\deg_B(\det\Delta_{x_0,j}(\tau_0+\delta,\cdot),\Omega), \tag{3.4}$$

where $\deg_B$ is the Brouwer degree. Let

$$\epsilon(x_0)=(-1)^n sign\det D\hat{F}(x_0). \tag{3.5}$$

We have the following global symmetric Hopf bifurcation theorem due to [24].

LEMMA 3.1. *Assume that* (A1)–(A4) *are satisfied and* $c_j(x_0,\tau_0,\beta_0)\neq 0$ *for some integer* $j\geq 0$ *and some* $(\tau_0,x_0,\beta_0)\in(0,\infty)\times M^\Gamma\times(0,\infty)$. *Let* S_j *denote the closure in* $[0,\infty)\times C(\mathbb{R};\mathbb{R}^n)\times[0,\infty)$ *of the set of all* $(\tau,z,\beta)\in[0,\infty)\times C(\mathbb{R};\mathbb{R}^n)\times\mathbb{R}\setminus M^*$ *such that* $x(t):=z(\frac{\beta}{2\pi}t)$ *is a* $\frac{2\pi}{\beta}$*-periodic solution of* (3.1) *with* $\rho x(t)=x(t-\frac{2\pi}{\beta}\frac{j}{N})$ *for* $t\in\mathbb{R}$. *Then* $S_j\neq\emptyset$, *and, for every bounded connected component* E_j *of* S_j, $(\Gamma\times S^1)E_j\cap M^*$ *is finite and*

$$\sum_{(\tau,x,\beta)\in(\Gamma\times S^1)E_j\cap M^*}\epsilon(x)c_j(x,\tau,\beta)=0; \tag{3.6}$$

here a set $E\subset(0,\infty)\times C(\mathbb{R};\mathbb{R}^n)\times(0,\infty)$ *is bounded if*

$$\sup\left\{\frac{1}{\tau}+\tau+\frac{1}{\beta}+\beta+\sup_{t\in\mathbb{R}}|x(t)|;\quad(\tau,x,\beta)\in E\right\}<\infty.$$

We now begin to apply the above result to discuss the global continuation of wave solutions of system (1.1). We need the following assumptions.

(H2) $\sup_{y\in\mathbb{R}}|h'(y)|<1$.

(H3) $g'(x)>0$ for all $x\in\mathbb{R}$.

PROPOSITION 3.2. *Assume that* (H1)–(H3) *are satisfied. Then system* (1.1) *has no nonconstant* 1*-periodic solution.*

Proof. By way of contradiction, let x be a nonconstant periodic solution of system (1.1) with $x_i(t)=x_i(t-1)$ for all $t\in\mathbb{R}$ and $i=1,2,3$. Then we obtain a system of ordinary differential equations

$$\begin{cases}\frac{1}{\tau}\dot{x}_1(t)=-x_1(t)+h(x_1(t))+2g(x_1(t))-g(x_2(t))-g(x_3(t)),\\ \frac{1}{\tau}\dot{x}_2(t)=-x_2(t)+h(x_2(t))+2g(x_2(t))-g(x_1(t))-g(x_3(t)),\\ \frac{1}{\tau}\dot{x}_3(t)=-x_3(t)+h(x_3(t))+2g(x_3(t))-g(x_2(t))-g(x_1(t)).\end{cases} \tag{3.7}$$

Note that the above equation is exactly the model equation for the Hopfield net [20] of three identical neurons with self-feedback, and thus

$$\begin{aligned}&V(x_1,x_2,x_3)\\&=-\frac{1}{2}\sum_{1\leq i<j\leq 3}[g(x_i)-g(x_j)]^2+\sum_{k=1}^{3}\int_0^{x_k}[s-h(s)]g\prime(s)ds\\&=g(x_1)g(x_2)+g(x_2)g(x_3)+g(x_3)g(x_1)\\&\quad-g^2(x_1)-g^2(x_2)-g^2(x_3)\\&\quad+\int_0^{x_1}[s-h(s)]g\prime(s)ds+\int_0^{x_2}[s-h(s)]g\prime(s)ds+\int_0^{x_3}[s-h(s)]g\prime(s)ds\end{aligned}$$

is the so-called energy function. For such an energy function, we have

$$\begin{aligned}
&\dot{V}_{(15)}(x_1,x_2,x_3)\\
&= g\prime(x_1)\dot{x}_1[g(x_2)+g(x_3)-2g(x_1)+x_1-h(x_1)]\\
&\quad + g\prime(x_2)\dot{x}_2[g(x_1)+g(x_3)-2g(x_2)+x_2-h(x_2)]\\
&\quad + g\prime(x_3)\dot{x}_3[g(x_1)+g(x_2)-2g(x_3)+x_3-h(x_3)]\\
&= -\tau\sum_{i=1}^{3} g\prime(x_i)(\dot{x}_i)^2 \le 0
\end{aligned}$$

and

$$\dot{V}_{(15)}(x_1,x_2,x_3)=0 \text{ if and only if } \dot{x}_1=\dot{x}_2=\dot{x}_3=0.$$

The LaSalle invariance principle [27] then implies that every solution of (3.6) converges to an equilibrium as $t\to\infty$. In particular, every 1-periodic solution of (1.1) must be constant. This completes the proof. □

PROPOSITION 3.3. *Under assumptions* (H1)–(H3)*, system* (1.1) *has no nonconstant* 2*-periodic solution.*

Proof. Assume that $x(t)$ is a 2-periodic solution. Let $x_4(t)=x_1(t-1), x_5(t)=x_2(t-1)$, and $x_6(t)=x_3(t-1)$. Then we obtain

$$\begin{cases}
\epsilon\dot{x}_1 = -x_1+h(x_4)-g(x_5)-g(x_6)+2g(x_4),\\
\epsilon\dot{x}_2 = -x_2+h(x_5)-g(x_4)-g(x_6)+2g(x_5),\\
\epsilon\dot{x}_3 = -x_3+h(x_6)-g(x_4)-g(x_5)+2g(x_6),\\
\epsilon\dot{x}_4 = -x_4+h(x_1)-g(x_2)-g(x_3)+2g(x_1),\\
\epsilon\dot{x}_5 = -x_5+h(x_2)-g(x_1)-g(x_3)+2g(x_2),\\
\epsilon\dot{x}_6 = -x_6+h(x_3)-g(x_1)-g(x_2)+2g(x_3).
\end{cases}$$

Then

$$\begin{cases}
\frac{1}{\tau}[x_1-x_4]\prime = -[x_1-x_4]+[h(x_4)-h(x_1)]\\
\qquad\qquad + [g(x_2)-g(x_5)+g(x_3)-g(x_6)-2(g(x_1)-g(x_4))],\\
\frac{1}{\tau}[x_2-x_5]\prime = -[x_2-x_5]+[h(x_5)-h(x_2)]\\
\qquad\qquad + [g(x_1)-g(x_4)+g(x_3)-g(x_6)-2(g(x_2)-g(x_5))],\\
\frac{1}{\tau}[x_3-x_6]\prime = -[x_3-x_6]+[h(x_6)-h(x_3)]\\
\qquad\qquad + [g(x_1)-g(x_4)+g(x_2)-g(x_5)-2(g(x_3)-g(x_6))].
\end{cases}$$

Let D^+ denote the upper right Dini derivative; then

$$\begin{cases}
\frac{1}{\tau}D^+|x_1-x_4| \le -|x_1-x_4|-2|g(x_1)-g(x_4)|+|h(x_1)-h(x_4)|\\
\qquad\qquad + |g(x_2)-g(x_5)|+|g(x_3)-g(x_6)|,\\
\frac{1}{\tau}D^+|x_2-x_5| \le -|x_2-x_5|-2|g(x_2)-g(x_5)|+|h(x_2)-h(x_5)|\\
\qquad\qquad + |g(x_1)-g(x_4)|+|g(x_3)-g(x_6)|,\\
\frac{1}{\tau}D^+|x_3-x_6| \le -|x_3-x_6|-2|g(x_3)-g(x_6)|+|h(x_3)-h(x_6)|\\
\qquad\qquad + |g(x_1)-g(x_4)|+|g(x_2)-g(x_5)|.
\end{cases}$$

Therefore,

$$\begin{aligned}
&\frac{1}{\tau}D^+[|x_1-x_4|+|x_2-x_5|+|x_3-x_6|]\\
&\le -[|x_1-x_4|+|x_2-x_5|+|x_3-x_6|]\\
&\qquad +|h(x_1)-h(x_4)|+|h(x_2)-h(x_5)|+|h(x_3)-h(x_6)|\\
&\le -\left[1-\sup_{\theta\in\mathbb{R}}|h\prime(\theta)|\right][|x_1-x_4|+|x_2-x_5|+|x_3-x_6|].
\end{aligned}$$

This implies that

$$|x_1(t)-x_4(t)|+|x_2(t)-x_5(t)|+|x_3(t)-x_6(t)|\to 0 \quad \text{as } t\to\infty.$$

Therefore, for a 2-periodic solution x of (1), we must have $x_1(t)=x_1(t-1), x_2(t)=x_2(t-1)$, and $x_3(t)=x_3(t-1)$. So Proposition 3.2 can be applied to conclude that x must be constant. This completes the proof. □

It remains to obtain a priori bounds for the norm of periodic solutions of (1.1). We need the following assumption.

(H4) $\sup_{y\in\mathbb{R}}[|h(y)|+|g(y)|]<\infty$.

PROPOSITION 3.4. *Assume* (H1)–(H4) *are satisfied. Then there exists* $M=M(h,g)>0$ *such that* $|x_1(t)|+|x_2(t)|+|x_3(t)|\le M$ *for all* $t\in\mathbb{R}$ *and for every periodic solution* x *of* (1.1).

Proof. Let $t^*\in\mathbb{R}$ and $j\in\{1,2,3\}$ be given so that $|x_j(t^*)|=\max_{t\in\mathbb{R}}\max_{1\le i\le 3}|x_i(t)|$. Then $\dot{x}_j(t^*)=0$. That is,

$$x_j(t^*)=h(x_j(t^*-1))-[g(x_{j-1}(t^*-1))+g(x_{j+1}(t^*-1))-2g(x_j(t^*-1))],$$

from which it follows that

$$|x_j(t^*)|\le\sup_{y\in\mathbb{R}}|h(y)|+4\sup_{y\in\mathbb{R}}|g(y)|:=\frac{M}{3}<\infty.$$

This completes the proof. □

We now apply Lemma 3.1 to investigate the global continuation of standing, mirror-reflecting, and discrete waves.

First, note that near $\tau=\tau_k$ system (1.1) has two bifurcations of discrete waves satisfying $x_{i-1}(t)=x_i(t\pm\frac{\omega}{3})$, where ω is a period. To look at the global continuation of such local bifurcations, we regard system (1.1) as a functional differential equation equivariant with respect to the action of $\Gamma=Z_3$, where the action is the cyclic permutation. We have

$$\begin{aligned}
M^\Gamma&=\{x\in\mathbb{R}^3;\gamma x=x \text{ for } \gamma\in\Gamma, F(\bar{x})=0\}\\
&=\{x\in\mathbb{R}^3;x_1=x_2=x_3 \text{ and } x_1=h(x_1)\}=\{0\}
\end{aligned}$$

under assumption (H2). Clearly, (A1) and (A2) are satisfied.

Under assumption (H1), the discussions in the last section show that

$$M^*=\{(\tau_k,0,\beta_k);\quad k\ge 1\}.$$

Therefore, M^* is discrete in $\mathbb{R}^3$.

Using Proposition 2.4 (ii), for a fixed integer k, we can choose positive constants b,c, and δ so that the only possible eigenvalue $u+iv$ of $A(\tau_k)$ with $(u,v)\in\partial\Omega$ is $i\beta_k$,

where $\Omega = (0, b) \times (\beta_k - c, \beta_k + c)$, and if $(\tau, \beta) \in [\tau_k - \delta, \tau_k + \delta] \times [\beta_k - c, \beta_k + c]$, then $i\beta$ is an eigenvalue of $A(\tau)$ if and only if $\tau = \tau_k$ and $\beta = \beta_k$. Then, using Proposition 2.4 (i), we can conclude that the analytic function $p_\tau(\lambda) := \lambda + \tau - \gamma\tau e^{-\lambda}$ has no zero in $\bar{\Omega}$ for $\tau = \tau_k \pm \delta$. Also, by Propositions 2.4 and 2.6, the above b, c, and δ can be chosen so that, for the analytic function

$$q_\tau(\lambda) = \lambda + \tau - (\gamma + 3\beta)\tau e^{-\lambda},$$

we have that $q_{\tau_k - \delta}$ has no zero in $\bar{\Omega}$, while $q_{\tau_k + \delta}$ has exactly one zero in $\bar{\Omega}$, and this zero is simple and is in the interior of $\bar{\Omega}$. Therefore,

$$\deg_B(q_{\tau_k - \delta}, \Omega) = 0,$$

and

$$\deg_B(q_{\tau_k + \delta}, \Omega) = 1.$$

With respect to the complexification of the above ($\Gamma = Z_3$)-action in $\mathbb{R}^3$, we have the isotypical decomposition

$$\mathbb{C}^3 = \mathbb{C}^3_0 \oplus \mathbb{C}^3_1 \oplus \mathbb{C}^3_2,$$

where

$$\mathbb{C}^3_j = \{(1, e^{i\frac{2\pi}{3}j}, e^{i\frac{4\pi}{3}j})x; \quad x \in \mathbb{C}\}.$$

We have shown that

$$\begin{aligned} \Delta_{0,j} &:= \Delta_0(\tau, \lambda)|_{\mathbb{C}^3_j} = \Delta(\tau, \lambda)|_{\mathbb{C}^3_j} \\ &= \begin{cases} \lambda + \tau - \gamma\tau e^{-\lambda} & \text{if } j = 0, \\ \lambda + \tau - (\gamma + 3\beta)\tau e^{-\lambda} \text{ if} & j = 1, 2. \end{cases} \end{aligned}$$

Therefore, from the above discussions, we get

$$c_0(0, \tau_k, \beta_k) = \deg_B(p_{\tau_k - \delta}, \Omega) - \deg_B(p_{\tau_k + \delta}, \Omega) = 0,$$

and, for $j = 1, 2$,

$$c_j(0, \tau_k, \beta_k) = \deg_B(q_{\tau_k - \delta}, \Omega) - \deg_B(q_{\tau_k + \delta}, \Omega) = -1.$$

Let $S_j, j = 1, 2$, denote the closure in $[0, \infty) \times C(\mathbb{R}; \mathbb{R}^3) \times [0, \infty)$ of the set of all triples $(\tau, z, \beta) \notin M^*$ such that $x(t) := z(\frac{\beta}{2\pi}t)$ is a $\frac{2\pi}{\beta}$-periodic solution of (1.1) with $x_{k+1}(t) = x_k(t - \frac{2\pi}{\beta}\frac{j}{3})$ for $t \in \mathbb{R}$ and $k = 1, 2, 3 (\mathrm{mod}\, 3)$. Then Lemma 3.1 implies that S_j must have a nonempty connected component E_j passing through $(\tau_k, 0, \beta_k)$, and this component must be unbounded in the sense that

$$\sup_{(\tau, x, \beta) \in E_j} \left\{ \tau + \frac{1}{\tau} + \beta + \frac{1}{\beta} + \sup_{t \in \mathbb{R}} |z(t)| \right\} = \infty,$$

for otherwise, the summation (3.6) must hold, and this is clearly impossible as $c_j(0, \tau_k, \beta_k)$ has the same sign for all positive integers k.

The projection of E_j onto the space $C(\mathbb{R};\mathbb{R}^3)$ is bounded due to Proposition 3.4. Near τ_k, (ii) of Proposition 2.4 shows that, for $(\tau,z,\beta)\in E_j$, we have

$$\frac{2\pi}{\beta}\in\left(\frac{2\pi}{2k\pi},\frac{2\pi}{2k\pi-\frac{\pi}{2}}\right)\subset\left(\frac{1}{k},\frac{1}{k-\frac{1}{4}}\right)\subset\left(\frac{1}{k},\frac{4}{3}\right)\subset\left(\frac{1}{k},2\right).$$

On the other hand, Propositions 3.2 and 3.3 imply that the projection of E_j onto the β-plane can never reach the lines $\frac{2\pi}{\beta}=\frac{1}{k}$ (note that (1.1) has no $\frac{1}{k}$-periodic solution as it does not have a 1-periodic solution) and $\frac{2\pi}{\beta}=2$. Therefore, the projection of E_j onto the β-plane always satisfies $\pi<\beta<2k\pi$.

On the other hand, the result of [28] shows that there exists $\alpha^*>0$ such that any period p of a periodic solution of (1.2) must satisfy $p\geq\alpha^*$. Consequently, for $(\tau,z,\beta)\in E_j$, we must have $\tau\frac{2\pi}{\beta}\geq\alpha^*$. That is, $\tau\geq\frac{\beta\alpha^*}{2\pi}>\frac{\alpha^*}{2}$ for every $\tau\in I$, the projection of E_j onto the τ-axis which must be an interval. Therefore, I must be unbounded from above. Clearly, I contains τ_k. This proves the following.

THEOREM 3.5. *For each $\tau>\tau_k$, system (1.1) always has two discrete waves satisfying $x_{j+1}(t)=x_j(t\pm\frac{\omega}{3})$ for $t\in\mathbb{R}$ and $j(\mathrm{mod}\,3)$, where ω is a period of $x(t)$ and $\frac{1}{k}<\omega<2$.*

Let us now consider the global continuation of mirror-reflecting waves and standing waves. For this purpose, we consider (1.1) as a functional differential equation equivariant with respect to the action of $\Gamma=Z_2$ on $\mathbb{R}^3$ defined by

$$\rho\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}x_1\\x_3\\x_2\end{pmatrix},\quad x_i\in\mathbb{R},i=1,2,3,Z_2=\langle\rho\rangle.$$

In this case,

$$M^\Gamma=\{x\in\mathbb{R}^3;\quad x_2=x_3,x_i=h(x_i)-g(x_{i-1})-g(x_{i+1})+2g(x_i),i(\mathrm{mod}\,3)\}.$$

The structure of M^Γ is explicitly described in the following proposition under the following assumption.

(H5) $yh''(y)<0$ and $yg''(y)<0$ for $y\neq0$.

PROPOSITION 3.6. *Under* (H1)–(H5)*, the system of equations*

$$x_i=h(x_i)-g(x_{i-1})-g(x_{i+1})+2g(x_i),\quad i(\mathrm{mod}\,3),\tag{3.8}$$

and

$$x_2=x_3\tag{3.9}$$

for $x=(x_1,x_2,x_3)^T$ has exactly three solutions. They are

$$(0,0,0)^T,\quad(z^-,y^+,y^+)^T,\quad(z^+,y^-,y^-)^T,$$

where $y^+>0,y^-<0,z^+>0,z^-<0$ are the unique solutions of

$$\begin{cases}y^\pm-h(y^\pm)=u^\pm,\\ z^\mp-h(z^\mp)=-2u^\pm\end{cases}\tag{3.10}$$

and $u^+>0$ and $u^-<0$ are the unique positive and negative solutions of

$$u+g[G^{-1}(-2u)]-g[G^{-1}(u)]=0\tag{3.11}$$

with $G : \mathbb{R} \to \mathbb{R}$ being given by the equation

$$G(\theta) = \theta - h(\theta), \quad \theta \in \mathbb{R}. \tag{3.12}$$

In other words,

$$M^\Gamma = \{(0,0,0)^T, (z^-, y^+, y^+)^T, (z^+, y^-, y^-)^T\}.$$

Proof. Under assumption (H2), $G : \mathbb{R} \to \mathbb{R}$ defined by (3.12) is an increasing function. Define

$$u = G(y), \quad v = G(z). \tag{3.13}$$

Then $x = (x_1, x_2, x_3)^T$ with $x_1 = z$ and $x_2 = x_3 = y$ satisfies (3.8) if and only if

$$u = g[G^{-1}(u)] - g[G^{-1}(v)] \tag{3.14}$$

and

$$v = -2u. \tag{3.15}$$

In other words, (u, v) is given by $v = -2u$ and $u = g[G^{-1}(u)] - g[G^{-1}(-2u)]$. Let

$$H(u) = u + g[G^{-1}(-2u)] - g[G^{-1}(u)], \quad u \in \mathbb{R}.$$

Then

$$H(0) = 0, \quad H(\pm\infty) = \pm\infty.$$

Note that

$$\begin{aligned} H'(u) &= 1 + g'[G^{-1}(-2u)](G^{-1})'(-2u)(-2) - g'[G^{-1}(u)](G^{-1})'(u) \\ &= 1 - 2g'[G^{-1}(-2u)](G^{-1})'(-2u) - g'[G^{-1}(u)](G^{-1})'(u). \end{aligned}$$

Implicitly differentiating $F(\theta) = \theta - h(\theta)$, we get

$$(G^{-1})'(\theta) = \frac{1}{1 - h'[G^{-1}(\theta)]}.$$

Therefore,

$$H'(u) = 1 - \frac{2g'[G^{-1}(-2u)]}{1 - h'[G^{-1}(-2u)]} - \frac{g'[G^{-1}(u)]}{1 - h'[G^{-1}(u)]}.$$

In particular, with $h'(0) = \gamma$ and $g'(0) = \beta$ and under assumption (H1), we have

$$H'(0) = 1 - \frac{2\beta}{1-\gamma} - \frac{\beta}{1-\gamma} = \frac{1 - (\gamma + 3\beta)}{1-\gamma} < 0.$$

Therefore, there must be $u^+ > 0$ and $u^- < 0$ such that $H(u^\pm) = 0$.

It remains to show that there exists no other nonzero zero of H. By way of contradiction, if there exists $u^* > 0$ (the case in which $u^* < 0$ can be dealt with

similarly) such that $H(u^*) = 0$ and $u^* \neq u^+$, then there must be $\theta > 0$ so that $H''(\theta) = 0$. However, we have

$$\begin{aligned} H''(u) &= -2g''[G^{-1}(-2u)][(G^{-1})'(-2u)]^2(-2) \\ &\quad -2g'[G^{-1}(-2u)](G^{-1})''(-2u)(-2) \\ &\quad -g''[G^{-1}(u)][(G^{-1})'(u)]^2 - g'[G^{-1}(u)](G^{-1})''(u) \\ &= 4g''[G^{-1}(-2u)][(G^{-1})'(-2u)]^2 + 4g'[G^{-1}(-2u)](G^{-1})''(-2u) \\ &\quad -g''[G^{-1}(u)][(G^{-1})'(u)]^2 - g'[G^{-1}(u)](G^{-1})''(u). \end{aligned}$$

Under assumption (H5), for $u > 0$ we have

$$g''[G^{-1}(-2u)] > 0, \quad g''[G^{-1}(u)] < 0.$$

Therefore, $H''(u) > 0$ if we can show that

$$(G^{-1})''(-2u) > 0 \text{ and } (G^{-1})''(u) < 0 \text{ for } u > 0. \tag{3.16}$$

The above holds by using (H5) since

$$(G^{-1})''(u) = \frac{h''(G^{-1}(u))(G^{-1})'(u)}{[1 - h'(G^{-1}(u))]^2}$$

has the opposite sign from u. (Recall that $G^{-1}(u)$ has the same sign as u.)

This completes the proof. □

To verify (A2) and (A4) in the case in which $\Gamma = Z_2$, we need the following condition.

(H6) $h'(\alpha) > 0, h'(\alpha) + 3g'(\alpha) < 1$, where $\alpha = y^{\pm}, z^{\pm}$.

The linearization of (1.1) at (z^*, y^*, y^*) with $z^* = z^{\mp}, y^* = y^{\pm}$ takes the form

$$\begin{cases} \frac{1}{\tau}\dot{X}_1(t) = -X_1(t) + h_1'(z^*)X_1(t-1) \\ \qquad\qquad -[g'(y^*)X_2(t-1) + g'(y^*)X_3(t-1) - 2g'(z^*)X_1(t-1)], \\ \frac{1}{\tau}\dot{X}_2(t) = -X_2(t) + h_1'(y^*)X_2(t-1) \\ \qquad\qquad -[g'(y^*)X_3(t-1) + g'(z^*)X_1(t-1) - 2g'(y^*)X_2(t-1)], \\ \frac{1}{\tau}\dot{X}_3(t) = -X_3(t) + h_1'(y^*)X_3(t-1) \\ \qquad\qquad -[g'(z^*)X_1(t-1) + g'(y^*)X_2(t-1) - 2g'(y^*)X_3(t-1)], \end{cases}$$

and the characteristic matrix becomes

$$\begin{aligned} &\Delta_{(z^*,y^*,y^*)}(\tau,\lambda) \\ &= \begin{pmatrix} A & \tau g'(y^*)e^{-\lambda} & \tau g'(y^*)e^{-\lambda} \\ \tau g'(z^*)e^{-\lambda} & B & \tau g'(y^*)e^{-\lambda} \\ \tau g'(z^*)e^{-\lambda} & \tau g'(y^*)e^{-\lambda} & B \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} A &= \lambda + \tau - \tau[h'(z^*) + 2g'(z^*)]e^{-\lambda}, \\ B &= \lambda + \tau - \tau[h'(y^*) + 2g'(y^*)]e^{-\lambda}. \end{aligned}$$

The isotypical decomposition of $\mathbb{C}^3$ with respect to the above $\Gamma = Z_2$ action is

$$\mathbb{C}^3 = \mathbb{C}_0^3 \oplus \mathbb{C}_1^3,$$

where

$$\mathbb{C}_0^3 = \{(x, y, y)^T; x, y \in \mathbb{C}\},$$
$$\mathbb{C}_1^3 = \{(0, z, -z)^T; z \in \mathbb{C}\}.$$

Therefore,

$$\Delta_{(z^*,y^*,y^*)}(\tau,\lambda)|_{\mathbb{C}_0^3} = \begin{pmatrix} \lambda + \tau - \tau[h'(z^*) + 2g'(z^*)]e^{-\lambda} & \tau g'(z^*)e^{-\lambda} \\ 2\tau g'(y^*)e^{-\lambda} & \lambda + \tau - \tau[h'(y^*) + g'(y^*)]e^{-\lambda} \end{pmatrix}$$

and

$$\Delta_{(z^*,y^*,y^*)}(\tau,\lambda)|_{\mathbb{C}_1^3} = \lambda + \tau - \tau[h'(y^*) + 3g'(y^*)]e^{-\lambda\tau}.$$

It is already shown in the proof of Proposition 2.4 (i) that, under assumption (H6), every zero of $\Delta_{(z^*,y^*,y^*)}(\tau,\lambda)|_{\mathbb{C}_1^3}$ has negative real part. Note that $\Delta_{(z^*,y^*,y^*)}(\tau,\lambda)|_{\mathbb{C}_0^3}$ is the characteristic matrix for the following linear system of delay differential equations:

$$\begin{cases} \frac{1}{\tau}\dot{u}_1(t) = -u_1(t) + [h'(z^*) + 2g'(z^*)]u_1(t-1) - g'(z^*)u_2(t-1), \\ \frac{1}{\tau}\dot{u}_2(t) = -u_2(t) + [h'(y^*) + g'(y^*)]u_2(t-1) - 2g'(y^*)u_1(t-1). \end{cases} \tag{3.17}$$

Let $V(u_1, u_2) = \max\{|u_1|, |u_2|\}$. For a given solution of (3.17), if at some $t \geq 0$ we have $V(u_1(t-1), u_2(t-1)) \leq V(u_1(t), u_2(t)) = |u_1(t)|$, then

$$\begin{aligned} &\frac{1}{\tau}D^+V(u_1(t), u_2(t)) \\ &\leq -|u_1(t)| + [h'(z^*) + 2g'(z^*)]|u_1(t-1)| + g'(z^*)|u_2(t-1)| \\ &\leq -|u_1(t)| + [h'(z^*) + 3g'(z^*)]|u_1(t)| \\ &= -[1 - h'(z^*) - 3g'(z^*)]V(u_1(t), u_2(t)). \end{aligned}$$

Similarly, for a given solution of (3.17), if at some $t \geq 0$ we have $V(u_1(t-1), u_2(t-1)) \leq V(u_1(t), u_2(t)) = |u_2(t)|$, then

$$\frac{1}{\tau}D^+V(u_1(t), u_2(t)) \leq -[1 - h'(y^*) - 3g'(y^*)]V(u_1(t), u_2(t)).$$

Therefore, using assumption (H6) and the Razumikhin-type LaSalle invariance principle in [27, 29], we can conclude that all solutions of (3.17) converge to zero as $t \to \infty$. This shows that all zeros of $\det\Delta_{(z^*,y^*,y^*)}(\tau,\lambda)|_{\mathbb{C}_0^3}$ have negative real parts. In particular, $\det\Delta_{(z^*,y^*,y^*)}(\tau,0)|_{\mathbb{C}_0^3} \neq 0$, and this determinant is exactly the determinant of the derivative of the corresponding F at (z^*, y^*, y^*). This shows that (A2) is satisfied and that (A3) is trivial.

Therefore, even in the case in which $\Gamma = Z_2$, we have

$$M^* = \{(\tau_k, 0, \beta_k); \quad k \geq 1\}.$$

Thus M^* is discrete and (A4) holds. Using similar arguments as for Theorem 3.5, we can get the following theorems.

THEOREM 3.7. *For each $\tau > \tau_k, k \geq 1$, system (1.1) has one standing wave satisfying $x_1(t) = x_1(t - \frac{\omega}{2})$ and $x_2(t) = x_3(t - \frac{\omega}{2})$ for $t \in \mathbb{R}$, where ω is a period of x and $\frac{1}{k} < \omega < 2$.*

THEOREM 3.8. *For each $\tau > \tau_k, k \geq 1$, system (1.1) has one mirror-reflecting wave satisfying $x_2(t) = x_3(t)$ and $x_i(t) = x_i(t + \omega)$ for $t \in \mathbb{R}, i = 1, 2, 3$, where $\frac{1}{k} < \omega < 2$.*

Remark 1. Due to the D_3-symmetry, Theorems 3.5–3.8 in fact imply the existence of three standing waves, three mirror-reflecting waves, and two discrete waves for each $\tau > \tau_k$. Note also that

$$\tau_1 < \tau_2 < \tau_3 < \cdots .$$

The above results establish the existence of $3k$ standing waves, $3k$ mirror-reflecting waves, and $2k$ discrete waves. It should be mentioned that, in the above theorems, ω is not necessarily the minimal period, and several branches of waves may coincide at some values of τ. In terms of the following five remarks, we can claim that for $\tau > \tau_1$, system (1.1) has three orbits of waves—one orbit of discrete waves, one orbit of standing waves, and one orbit of mirror-reflecting waves—and only the last two orbits may coincide through the mechanism of periodic doubling. Discounting the above possible coincidence, system (1.1) has at least five wave solutions for each $\tau > \tau_1$.

Remark 2. A branch of nontrivial discrete waves and a branch of mirror-reflecting waves cannot coincide at any value of τ, for otherwise there exists a nontrivial ω-periodic solution x of (1.1) such that $x_i(t) = x_{i-1}(t \pm \frac{\omega}{3})$ for $i (\mathrm{mod}\, 3)$ and $x_j(t) = x_k(t)$ for some $j \neq k$. For simplicity, let $x_2(t) = x_3(t)$. Then $x_2(t) = x_3(t \pm \frac{\omega}{3})$ implies that $\frac{\omega}{3}$ is also a period of $x_2 = x_3$, and thus $x_1(t) = x_2(t \pm \frac{\omega}{3}) = x_2(t) (= x_3(t))$. So x must be spatially homogeneous. As $\sup_{x \in \mathbb{R}} |h'(x)| < 1$ implies that $y = 0$ is the global attractor of the scalar equation $y'(t) = -y(t) + h(y(t - \tau))$ for any $\tau \geq 0$ (see, for example, [16]), we have $x = 0$, which is a contradiction.

Remark 3. A branch of nontrivial discrete waves and a branch of standing waves cannot coincide at any value of τ, for otherwise there exists a nontrivial ω-periodic solution x of (1.1) such that $x_i(t) = x_{i-1}(t \pm \frac{\omega}{3})$ for $i (\mathrm{mod}\ 3)$ and, say, $x_1(t) = x_1(t + \frac{\omega}{2}), x_2(t) = x_3(t + \frac{\omega}{2})$. Then $x_2(t) = x_3(t + \frac{\omega}{3}) = x_3(t + \frac{\omega}{2})$. (The other case in which $x_2(t) = x_3(t - \frac{\omega}{3})$ can be dealt similarly.) Therefore, $\frac{\omega}{6}$ is also a period of x_3 (and thus x_2). Consequently, $x_2(t) = x_3(t + \frac{\omega}{3}) = x_3(t)$ and $x_1(t) = x_2(t + \frac{\omega}{3}) = x_2(t) = x_3(t)$. Again, x must be spatially homogeneous, and thus $x = 0$, which is a contradiction.

Remark 4. A branch of nontrivial discrete waves of the form $x_i(t) = x_{i-1}(t - \frac{\omega}{3})$ and a branch of discrete waves of the form $x_i(t) = x_{i-1}(t + \frac{\omega}{3})$ for $i (\mathrm{mod}\, 3)$ and $t \in \mathbb{R}$ cannot coincide at any value of τ. Again, this can be verified by way of contradiction. Namely, if there is a discrete wave satisfying simultaneously $x_i(t) = x_{i-1}(t + \frac{\omega}{3}) = x_{i-1}(t - \frac{\omega}{3})$ for $i (\mathrm{mod}\, 3)$, then $\frac{2\omega}{3}$ and ω are periods of x, and so is $\frac{\omega}{3}$. This, together with $x_i(t) = x_{i-1}(t - \frac{\omega}{3})$, implies that x is spatially homogeneous, and thus $x = 0$, which is a contradiction.

Remark 5. As no nontrivial spatially homogeneous periodic solution exists, it is clear that a branch of nontrivial mirror-reflecting waves satisfying $x_i(t) = x_j(t)$ for some $i \neq j$ and a branch of mirror-reflecting waves satisfying $x_l(t) = x_m(t)$ for some $l \neq m$ cannot coincide at any value of τ if $(i, j) \neq (l, m)$. Similarly, a branch of nontrivial standing waves with $x_i(t) = x_i(t + \frac{\omega}{2}), x_j(t) = x_k(t + \frac{\omega}{2})$ for $i \neq j \neq k$ and a branch of nontrivial standing waves with $x_{i^*}(t) = x_{j^*}(t + \frac{\omega}{2}), x_{j^*}(t) = x_{k^*}(t + \frac{\omega}{2})$ for $i^* \neq j^* \neq k^*$ cannot coincide at any value of τ if $(i, j, k) \neq (i^*, j^*, k^*)$.

Remark 6. Unfortunately, the above arguments cannot be extended to rule out the possibility of the coincidence of a branch of nontrivial ω-periodic mirror-reflecting waves with $x_i(t) = x_j(t)$ for some $i \neq j$ and a branch of ω-periodic standing waves with $x_i(t) = x_j(t + \frac{\omega}{2})$ for some $i \neq j$. In fact, such a coincidence may occur at some value of τ where periodic doubling happens: $x_i(t) = x_i(t + \frac{\omega}{2}), i(\mathrm{mod}\, 3), t \in \mathbb{R}$.

Acknowledgment. We wish to thank one reviewer for her or his valuable comments that led to truly significant improvement of the manuscript.

REFERENCES

[1] J. Szenfagothai, *The "module-concept" in cerebral cortex architecture*, Brain Research, 95 (1967), pp. 475–496.
[2] J. C. Eccles, M. Ito, and J. Szenfagothai, *The Cerebellum as Neuronal Machine*, Springer-Verlag, New York, 1967.
[3] P. Anderson, O. Gross, T. Lomo, and O. Sveen, *Participation of inhibitory interneurons in the control of hippocampal cortical output*, in The Interneuron, M. Brazier, ed., University of California Press, Los Angeles, 1969.
[4] F. C. Hoppensteadt, *An Introduction to the Mathematics of Neurons*, Cambridge University Press, New York, 1986.
[5] G. M. Shepherd, *The Synaptic Organization of the Brain*, Oxford University Press, New York, 1990.
[6] R. J. Douglas, A. C. Martin, and D. Whitteridge, *A canonical microcircuit for neocortex*, Neural Comp., 1 (1989), pp. 480–488.
[7] J. Milton, *Dynamics of Small Neural Populations*, AMS, Providence, RI, 1996.
[8] J. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.
[9] J. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.
[10] O. Diekmann, S. A. van Gils, S. M. Verduyn Lunel, and H.-O. Walther, *Delay Equations. Functional, Complex, and Nonlinear Analysis*, Springer-Verlag, New York, 1995.
[11] Y. Chen and J. Wu, *Existence and attraction of a phase-locked oscillation in a delayed network of two neurons*, Differential Integral Equations, 14 (2001), pp. 1181–1236.
[12] Y. Chen and J. Wu, *Minimal instability and unstable set of a phase-locked orbit in a delayed neural network*, Phys. D, 134 (1999), pp. 185–199.
[13] Y. Chen, J. Wu, and T. Krisztin, *Connecting orbits from synchronous periodic solutions to phase-locked periodic solutions in a delay differential system*, J. Differential Equations, 163 (2000), pp. 130–173.
[14] J. Wu, *Introduction to Neural Dynamics and Signal Transmission Delay*, Walter de Gruyter, Berlin, 2001.
[15] J. Mallet-Paret and G. Sell, *Systems of differential delay equations: Floquet multipliers and discrete Lyapunov functions*, J. Differential Equations, 125 (1996), pp. 380–440.
[16] J. Mallet-Paret and G. Sell, *The Poincaré-Bendixson theorem for monotone cyclic feedback systems with delay*, J. Differential Equations, 125 (1996), pp. 441–489.
[17] T. Krisztin. H.-O. Walther, and J. Wu, *Shape, Smoothness and Invariant Stratification of an Attracting Set for Delayed Monotone Positive Feedback*, AMS, Providence, RI, 1999.
[18] M. Golubitsky, I. Stewart, and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Springer-Verlag, New York, 1988.
[19] M. A. Cohen and S. Grossberg, *Absolute stability of global pattern formation and parallel memory storage by competitive neural networks*, IEEE Trans. Systems Man Cybernet., 13 (1983), pp. 815–826.
[20] J. J. Hopfield, *Neurons with graded response have collective computational properties like those of two-stage neurons*, Proc. Nat. Acad. Sci. USA, 81 (1984), pp. 3088–3092.
[21] C. M. Marcus and R. M. Westervelt, *Stability of analog neural networks with delay*, Phys. Rev. A, 39 (1989), pp. 347–359.
[22] J. Bélair, *Stability in a model of a delayed neural network*, J. Dynam. Differential Equations, 5 (1993), pp. 607–623.
[23] J. Wu, *Symmetric functional differential equations and neural networks with memory*, Trans. Amer. Math. Soc., 350 (1998), pp. 4799–4838.
[24] W. Krawcewicz and J. Wu, *Theory and applications of Hopf bifurcations in symmetric functional differential equations*, Nonlinear Anal., 35 (1999), pp. 845–870.

[25] W. Krawcewicz and J. Wu, *Theory of Degrees with Applications to Bifurcations and Differential Equations*, John Wiley and Sons, Boston, 1996.
[26] B. W. Levinger, *A folk theorem in functional differential equations*, J. Differential Equations, 4 (1968), pp. 612–619.
[27] J. LaSalle, *The Stability of Dynamical Systems*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 25, SIAM, Philadelphia, 1976.
[28] A. Lasota and J. A. Yorke, *Bounds for periodic solutions of differential equations in Banach spaces*, J. Differential Equations, 10 (1971), pp. 83–91.
[29] J. R. Haddock and J. Terjéki, *Liapunov-Razumikhin functions and invariance principle for functional differential equations*, J. Differential Equations, 48 (1983), pp. 95–122.

SIAM J. MATH. ANAL.
Vol. 34, No. 4, pp. 861–881

DYNAMIC BOUNDARY CONDITIONS FOR HAMILTON–JACOBI EQUATIONS*

C. M. ELLIOTT†, Y. GIGA‡, AND S. GOTO§

Abstract. A nonstandard dynamic boundary condition for a Hamilton–Jacobi equation in one space dimension is studied in the context of viscosity solutions. A comparison principle, and hence uniqueness, is proved by consideration of an equivalent notion of viscosity solution for an alternative formulation of the boundary condition. The relationship with a Neumann condition is established. Global existence is obtained by consideration of a related parabolic approximation with a dynamic boundary condition. The problem is motivated by applications in superconductivity and interface evolution.

Key words. Hamilton–Jacobi equation, dynamic boundary condition, viscosity solution

AMS subject classification. 35L60

PII. S003614100139957X

1. Introduction.

We consider the first order equation

$$u_t - F\left(u_x^2 + \gamma^2\right)^{1/2} = 0 \quad \text{in } \Omega \times (0,\infty) \tag{1.1}$$

supplemented with the dynamic boundary condition

$$u_t - F\alpha = 0 \quad \text{on } \partial\Omega \times (0,\infty), \tag{1.2}$$

where Ω is a bounded open interval. The functions F and α are given continuous functions on $\overline{\Omega} \times [0,\infty)$, $\partial\Omega \times [0,\infty)$, respectively, and $\gamma \geq 0$ is a constant.

There are at least two sources of this problem. Consider the mean field vortex density model in a cylinder $D \times \mathbb{R}$ ($D \subset \mathbb{R}^2$) when the magnetic field $\vec{H}$ is orthogonal to the axis of the cylinder; see Chapman [3]. The vorticity field $\vec{\omega} = (\nabla^\perp \psi, 0)$, $\nabla^\perp = (-\partial_{x_2}, \partial_{x_1})$ is required to satisfy the conservation of vorticity

$$\vec{\omega}_t + \operatorname{curl}(\vec{\omega} \times \vec{v}) = 0.$$

If the velocity field $\vec{v}$ is of the form

$$\vec{v} = \operatorname{curl}\vec{H} \times \vec{\omega}/|\vec{\omega}|$$

and $\vec{H}$ is given, then the conservation of vorticity yields

$$\psi_t = |\nabla\psi| F,$$

*Received by the editors December 12, 2001; accepted for publication (in revised form) July 31, 2002; published electronically February 25, 2003.
http://www.siam.org/journals/sima/34-4/39957.html

†School of Mathematical Sciences, University of Sussex, Falmer, Brighton BN1 9QH, Great Britain (C.M.Elliott@sussex.ac.uk).

‡Department of Mathematics, Hokkaido University, Sapporo 060-0810, Japan (giga@math.sci.hokudai.ac.jp). The work of this author was partially supported by Grant-in-Aid for Scientific Research 10304010, 12874024, the Japan Society of the Promotion of Science.

§Department of Computational Science, Faculty of Science, Kanazawa University, Kanazawa 920-1192, Japan (goto@kenroku.kanazawa-u.ac.jp). The work of this author was partially supported by Grant-in-Aid for Encouragement of Young Scientists 11740108, the Japan Society of the Promotion of Science.

where F is a given function. Our equation (1.1) is derived by assuming that $\partial_{x_2}\psi = \gamma$ is a constant on $D = \Omega \times \mathbb{R}$ if we set $u(x_1, t) = \psi(x_1, x_2, t) - \gamma x_2$. The quantity $-\psi_t$ on the boundary corresponds to the flux $\vec{n} \times (\vec{\omega} \times \vec{v})$ on $\partial D \times \mathbb{R}$. The condition $\psi_t = F\alpha$ is considered as a special case of assigning the value of flux, and we obtain (1.1), (1.2). A full system with a different boundary condition $\vec{\omega} \cdot \vec{n} = 0$ is studied by Elliott, Schätzle, and Stoth [6].

Another source of the problem is a surface evolution problem with dynamic boundary condition. Consider (1.1) with $\gamma = 1$. Then (1.1) is equivalent to requiring that the upward normal velocity V of the graph $\Gamma_t = \{y = u(x, t)\}$ equal F, i.e., $V = F$. The boundary condition (1.2) is equivalent to saying that the upward velocity v of Γ_t on $\partial\Omega \times \mathbb{R}$ is equal to $F\alpha$, i.e., $v = F\alpha$. In [1] Angenent and Gurtin derive a dynamic boundary condition for the mean curvature flow equation. It is of the form $v = A\cos\theta + B$, where v is the normal velocity of $\partial\Gamma_t$ in $\partial\Omega$ and θ is the contact angle of Γ_t and $\partial\Omega$; A and B are constants. Our boundary condition corresponds to the case $A = 0$.

Our goal is to study the unique global-in-time solvability of (1.1), (1.2) for a given initial data. Since the problem is of first order, it is convenient to handle this problem in the realm of viscosity solutions; see, e.g., Barles [2]. We establish the comparison principle (section 3) for (1.1) and (1.2) by deriving an equivalent definition (section 2) of solutions. Although the dynamic boundary value problem is studied in [2, p. 102, (4.23)], it is essentially of Neumann type and does not include (1.2). We further prove (section 5) that the solution of (1.1) and (1.2) solves the Neumann problem for (1.1) with

$$\partial u/\partial\nu = (\mathrm{Sign}F)\left\{(\alpha - \gamma)_+(\alpha + \gamma)\right\}^{1/2} \tag{1.3}$$

in the viscosity sense, where β_+ denotes the positive part of β and $\mathrm{Sign}F$ denotes the sign of F, i.e., $\mathrm{Sign}F = \pm 1$ if $F \gtrless 0$ and $\mathrm{Sign}F = 0$ if $F = 0$. It might be possible to prove the comparison principle for (1.1) with the inhomogeneous data $\partial u/\partial\nu = p(t)$ when p is continuous; see Claisse [4]. However, our comparison principle for (1.1) and (1.2) still holds when F changes sign, in which case the Neumann data in (1.3) is discontinuous and hence is not included in the literature. Moreover, our proof is more direct and does not use (1.3). Our comparison principle yields the uniqueness of viscosity solutions for (1.1) and (1.2).

We also prove the global existence (section 4) of a solution for (1.1), (1.2) when the initial data a is a Lipschitz function in $\overline{\Omega}$ by using the approximate equation

$$u_t - \varepsilon u_{xx} - F\left(u_x^2 + \gamma^2\right)^{1/2} = 0 \quad \text{in } \Omega \times (0, \infty) \tag{1.4}$$

with the dynamic boundary condition

$$u_t - F\alpha + \varepsilon\partial u/\partial\nu = 0 \quad \text{on } \partial\Omega \times (0, \infty), \tag{1.5}$$

where ε is a positive parameter. The dynamic boundary condition for uniformly parabolic equations is well studied, for example, by Hinterman [10] and Escher [7, 8]. Their results may be applied to (1.4) and (1.5) in order to yield at least a local solution. However, since the global existence of solutions is easy to show, we give a proof for global solvability of (1.4), (1.5) in the appendix. By the maximum principle we derive a priori bounds (section 4) for the sup norms of u_t^ε, u_x^ε, u^ε in $\overline{\Omega} \times [0, T]$ for the solutions of (1.4), (1.5) independent of $\varepsilon \in (0, 1)$. This yields the solution of (1.1), (1.2) as a limit as $\varepsilon \to 0$. The presence of the term $\varepsilon\partial u/\partial\nu$ in (1.5) is crucial in order to obtain the a priori bound.

Finally, we warn the reader that the boundary condition (1.2) cannot be replaced by a formally equivalent Dirichlet boundary condition

$$u(x,t) = \int_0^t F(x,\tau)\alpha(x,\tau)d\tau + a(x) \tag{1.6}$$

even in the viscosity sense. We give in section 5 an explicit solution of (1.1) which solves (1.2) (resp., (1.6)) but does not solve (1.6) (resp., (1.2)) when $\alpha \equiv 1$, $F \equiv 1$, and $\alpha > \gamma$.

2. Definitions and equivalent notions of solutions. Let Ω be a bounded interval $(0, L) \subset \mathbb{R}$ and let $T > 0$ be a constant. For brevity we set $Q = \Omega \times (0, T)$, $\hat{Q} = \overline{\Omega} \times (0, T)$ and their closure $\overline{Q} = \overline{\Omega} \times [0, T]$. Given a mapping $k := k(x, t, \tau, p) : \hat{Q} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ we recall the following definitions of viscosity sub- and supersolutions $u \in C(\hat{Q})$ for k.

DEFINITION 2.1. *A function u is said to be a viscosity subsolution of k (in $\hat{Q}$) provided for any $(\hat{x}, \hat{t}, \phi) \in \hat{Q} \times C^1(\hat{Q})$ such that*

$$(u - \phi)(\hat{x}, \hat{t}) = \sup_{\hat{Q}}(u - \phi);$$

then the inequality

$$k(\hat{x}, \hat{t}, \tau, p) \le 0$$

holds where $\tau = \phi_t(\hat{x}, \hat{t})$ and $p = \phi_x(\hat{x}, \hat{t})$.

DEFINITION 2.2. *A function u is said to be a viscosity supersolution of k (in $\hat{Q}$) provided for any $(\hat{x}, \hat{t}, \phi) \in \hat{Q} \times C^1(\hat{Q})$ such that*

$$(u - \phi)(\hat{x}, \hat{t}) = \inf_{\hat{Q}}(u - \phi);$$

then the inequality

$$k(\hat{x}, \hat{t}, \tau, p) \ge 0$$

holds where $\tau = \phi_t(\hat{x}, \hat{t})$ and $p = \phi_x(\hat{x}, \hat{t})$.

Let F and α be given functions in $C(\overline{Q})$, $C(\partial\Omega \times [0, T])$, respectively, and let $\gamma \ge 0$ be a given constant. We use the notation, since $\partial\Omega = \{0, L\}$, that $\frac{\partial}{\partial\nu} = \nu\frac{\partial}{\partial x}$ on $\partial\Omega$ with $\nu = -1$ for $x = 0$ and $\nu = +1$ for $x = L$. The initial boundary value problem is

$$\begin{cases} u_t - F\left(u_x^2 + \gamma^2\right)^{1/2} = 0 & \text{in} \quad Q, \\ u_t - F\alpha = 0 & \text{on} \quad \partial\Omega \times (0, T), \\ u|_{t=0} = a & \text{on} \quad \Omega. \end{cases} \tag{2.1}$$

In order to formulate the definition of a viscosity solution to (2.1) we define, for $(x, t, \tau, p) \in \hat{Q} \times \mathbb{R} \times \mathbb{R}$,

$$E(x,t,\tau,p) := \tau - F(x,t)\left(p^2 + \gamma^2\right)^{1/2},$$

$$F_{\min}(x,t,\tau,p) := \begin{cases} E(x,t,\tau,p) & \text{if} \quad x \in \Omega, \\ \min\{\tau - F(x,t)\alpha(x,t), E(x,t,\tau,p)\} & \text{if} \quad x \in \partial\Omega, \end{cases}$$

$$F_{\max}(x,t,\tau,p) := \begin{cases} E(x,t,\tau,p) & \text{if} \quad x \in \Omega, \\ \max\{\tau - F(x,t)\alpha(x,t), E(x,t,\tau,p)\} & \text{if} \quad x \in \partial\Omega. \end{cases}$$

DEFINITION 2.3. *We say that $u \in C(\overline{Q})$ is a viscosity solution of* (2.1) *provided $u(x,0) = a(x)$, $x \in \overline{\Omega}$, and u is a viscosity subsolution for $F_{\min}$ and a viscosity supersolution for $F_{\max}$.*

This is the usual notion of viscosity solution for boundary value problems (cf. [5]). We give an equivalent notion of solution by introducing, for $(x,t,\tau,p) \in \hat{Q} \times \mathbb{R} \times \mathbb{R}$,

$$G(x,t,\tau,p) := \begin{cases} E(x,t,\tau,p) & \text{if} \quad x \in \Omega, \\ \tau - F(x,t) \max\left\{\alpha(x,t), \left(([p\nu \mathrm{Sign} F]_-)^2 + \gamma^2\right)^{1/2}\right\} & \text{if} \quad x \in \partial\Omega, \end{cases}$$

where f_- is the negative part of f. Set

$$G_B(x,t,\tau,p) = \tau - F(x,t) \max\left\{\alpha(x,t), \left(([p\nu \mathrm{Sign} F]_-)^2 + \gamma^2\right)^{1/2}\right\}.$$

An alternative expression for G_B is

$$G_B(x,t,\tau,p) = \tau - F(x,t)\left([(\mathrm{Sign} F)p\nu - \delta][(\mathrm{Sign} F)p\nu + \delta]_- + \gamma^2 + \delta^2\right)^{1/2},$$

where $\delta = (\max(\alpha,\gamma)^2 - \gamma^2)^{1/2}$. This identity follows from

$$\max\left\{\alpha, ((\eta_-)^2 + \gamma^2)^{1/2}\right\} = \left((\eta - \delta)(\eta + \delta)_- + \gamma^2 + \delta^2\right)^{1/2}$$

for $\eta \in \mathbb{R}$, which is easy to prove. The main purpose of this section is to prove the following proposition.

PROPOSITION 2.4. *A function u is a viscosity solution of* (2.1) *if and only if $u \in C(\overline{Q})$, $u(x,0) = a(x)$, $x \in \Omega$, and u is both a viscosity subsolution and a viscosity supersolution for G.*

It is sufficient to prove the following lemmas.

LEMMA 2.5. *A function u is a viscosity supersolution for G if and only if u is a viscosity supersolution for $F_{\max}$.*

Proof. We use the notation for $(\hat{x},\hat{t},\phi,\tau,p)$ introduced in Definition 2.2, and $\hat{F} = F(\hat{x},\hat{t})$ and $\hat{\alpha} = \alpha(\hat{x},\hat{t})$. Clearly there is nothing to prove if $\hat{x} \in \Omega$ since G and $F_{\max}$ agree in Ω. Furthermore, without loss of generality we may assume $\hat{x} = L \in \partial\Omega$ so that $\nu = 1$. We suppress the word *viscosity* in the proof.

If u is a supersolution for G, then $\tau \geq \hat{F}\hat{\alpha}$ so that, trivially, u is a supersolution for $F_{\max}$. Thus the proof parts are concluding the situation that u is a supersolution for $F_{\max}$ and proving that this implies u is a supersolution for G. We have that

$$\max\left\{\tau - \hat{F}\hat{\alpha}, \tau - \hat{F}(p^2 + \gamma^2)^{1/2}\right\} \geq 0.$$

We may assume that $(L,\hat{t})$, by modifying ϕ if necessary, is a unique minimizer of $u - \phi$. It is convenient to make the following observation.

Observation 1. The following device shifts the minimizer into the interior. Let $h \in C^1(0,\infty)$ be a nonincreasing function such that $h(\sigma) = 0$ for all $\sigma \geqslant 1$, and that $h(\sigma) \to +\infty$ as $\sigma \to 0$. Set $d(x) := L - x$ and $\phi^\varepsilon(x,t) := \phi(x,t) - \varepsilon h(\frac{d(x)}{\varepsilon})$. Let $(x_\varepsilon, t_\varepsilon)$ be a minimum point of $u - \phi^\varepsilon$ on $\hat{Q}$. Since

$$\liminf_{\varepsilon\to 0 *}(u - \phi^\varepsilon) = u - \phi$$

on $\hat{Q}$, we see that $(x_\varepsilon, t_\varepsilon) \to (L,\hat{t})$ as $\varepsilon \to 0$. Here $\liminf_*$ is the relaxed limit as in [2], i.e.,

$$\liminf_{\varepsilon\to 0 *} f^\varepsilon(x,t) = \liminf_{\varepsilon\to 0}\{f^\delta(y,s); |y-x| < \varepsilon, |s-t| < \varepsilon, 0 < \delta < \varepsilon, (y,s) \in \hat{Q}\}.$$

Furthermore, since $h(\sigma) \to \infty$ as $\sigma \to 0$ we have that, for ε sufficiently small, $x_\varepsilon < L$ and $x_\varepsilon \in \Omega$. Because u is a supersolution for $F_{\max}$, this implies that

$$\phi_t^\varepsilon(x_\varepsilon, t_\varepsilon) \geq F(x_\varepsilon, t_\varepsilon) \left(\phi_x^\varepsilon(x_\varepsilon, t_\varepsilon)^2 + \gamma^2\right)^{1/2},$$

which yields

$$\phi_t(x_\varepsilon, t_\varepsilon) \geq F(x_\varepsilon, t_\varepsilon) \left(\left(\phi_x(x_\varepsilon, t_\varepsilon) + h'\left(\frac{d(x_\varepsilon)}{\varepsilon}\right)\right)^2 + \gamma^2\right)^{1/2}.$$

Observation 2. Set for $A > 0$, $\psi(x,t) := \phi(x,t) - Ad(x)$. Clearly

$$(u - \psi)(x,t) = (u - \phi)(x,t) + Ad(x) \geq (u - \psi)(L, \hat{t})$$

for all $(x,t) \in \hat{Q}$. Since u is a supersolution for $F_{\max}$, we have

$$\max\left\{\tau - \hat{F}\hat{\alpha}, \tau - \hat{F}\left((p + A)^2 + \gamma^2\right)^{1/2}\right\} \geq 0.$$

We consider separately the cases $\hat{F} > 0$ and $\hat{F} < 0$ since the case $\hat{F} = 0$ leads to $G = F_{\max}$.

Case I. $\hat{F} > 0$. Again we discuss three cases.

(i) $\gamma \geq \hat{\alpha}$ and $p \geq 0$. This is immediately treated by Observation 1. Sending ε to zero we have

$$\tau \geq \hat{F}\gamma = \hat{F} \max\left\{\hat{\alpha}, \left([p\nu]_-\right)^2 + \gamma^2\right)^{1/2}\right\}.$$

(ii) $p\nu + \delta < 0$. Using Observation 1 we set that for small ε, we have

$$\phi_x^\varepsilon(x_\varepsilon, t_\varepsilon) = \phi_x(x_\varepsilon, t_\varepsilon) + h'\left(\frac{d(x_\varepsilon)}{\varepsilon}\right) \leq \phi_x(x_\varepsilon, t_\varepsilon) < 0$$

so that

$$\phi_t(x_\varepsilon, t_\varepsilon) \geq F(x_\varepsilon, t_\varepsilon)(\phi_x(x_\varepsilon, t_\varepsilon)^2 + \gamma^2)^{1/2},$$

and sending ε to zero,

$$\tau \geq \hat{F}(p^2 + \gamma^2)^{1/2} = \hat{F}((p\nu - \delta)(p\nu + \delta)_- + \gamma^2 + \delta^2)^{1/2}.$$

(iii) $p\nu + \delta \geq 0$, $\gamma < \hat{\alpha}$. This is the remaining case. From Observation 2, by choosing A large we find

$$\tau \geq \hat{F}\hat{\alpha} = \hat{F}((p\nu - \delta)(p\nu + \delta)_- + \gamma^2 + \delta^2)^{1/2}.$$

Case II. $\hat{F} < 0$.

(i) $\gamma \geq \hat{\alpha}$, $p \geq 0$. This case leads to $F_{\max} = \tau - \hat{F}(p^2 + \gamma^2)^{1/2} = G_B$.

(ii) $\gamma \geq \hat{\alpha}$, $p < 0$. Use Observation 2 and set $A = -p > 0$, which yields $F_{\max} = \tau - \hat{F}\gamma = G_B$.

(iii) $\gamma < \hat{\alpha}$, $p \leq \delta$. If $|p| \leq \delta$, then $F_{\max} = \tau - \hat{F}\hat{\alpha} = G_B$. If $p < -\delta$, then $G_B = \tau - \hat{F}\hat{\alpha}$, and using Observation 2 and $A = -p$ we have $F_{\max} = \tau - \hat{F}\hat{\alpha}$.

(iv) $\gamma < \hat{\alpha}$, $p > \delta$. This case leads to $F_{\max} = \tau - \hat{F}(p^2 + \gamma^2)^{1/2} = G_B$. □

LEMMA 2.6. *A function u is a viscosity subsolution for G if and only if u is a viscosity subsolution for $F_{\min}$.*

Proof. We use the notation for $(\hat{x}, \hat{t}, \phi, \tau, p)$ introduced in Definition 2.1. Clearly there is nothing to prove if $\hat{x} \in \Omega$ since G and $F_{\min}$ agree in Ω. Furthermore, without loss of generality we may assume $\hat{x} = L \in \partial\Omega$ so that $\nu = 1$.

If u is a subsolution for G, then

$$\tau \le \hat{F} \max\left\{\hat{\alpha}, \left(([p\nu \mathrm{Sign}\hat{F}]_-)^2 + \gamma^2\right)^{1/2}\right\}$$

so that either $\tau \le \hat{F}\hat{\alpha}$ or $\tau \le \hat{F}(([p\nu\mathrm{Sign}\hat{F}]_-)^2 + \gamma^2)^{1/2}$. If $\hat{F} > 0$, then this implies $\min\{\tau - \hat{F}\hat{\alpha}, \tau - \hat{F}(p^2+\gamma^2)^{1/2}\} \le 0$ and u is a subsolution for $F_{\min}$.

On the other hand if $\hat{F} < 0$, then $\tau \le \hat{F}((p\nu+\delta)(p\nu-\delta)_+ + \gamma^2 + \delta^2)^{1/2}$. If $\gamma \ge \hat{\alpha}$, then either $\tau \le \hat{F}(p^2+\gamma^2)^{1/2}$ or $\tau \le \hat{F}\gamma \le \hat{F}\hat{\alpha}$, which implies u is a subsolution for $F_{\min}$. Whenever $\hat{\alpha} > \gamma$, then either $\tau \le \hat{F}(\gamma^2+\delta^2)^{1/2} = \hat{F}\hat{\alpha}$ or $\tau \le \hat{F}(p^2+\gamma^2)^{1/2}$ and again u is a subsolution for $F_{\min}$.

We may suppose that u is a subsolution for $F_{\min}$ so that

$$\min\left\{\tau - \hat{F}\hat{\alpha}, \tau - \hat{F}(p^2+\gamma^2)^{1/2}\right\} \le 0.$$

It is convenient to make the following observation.

Observation 1. The following device shifts the maximizer into the interior. Let $h \in C^1(0,\infty)$ be a nonincreasing function such that $h(\sigma) = 0$ for all $\sigma \ge 1$ and that $h(\sigma) \to +\infty$ as $\sigma \to 0$. Set $d(x) := L - x$ and $\phi^\varepsilon(x,t) := \phi(x,t) + \varepsilon h(\frac{d(x)}{\varepsilon})$. Let $(x_\varepsilon, t_\varepsilon)$ be a maximum point of $u - \phi^\varepsilon$ on $\hat{Q}$. Since

$$\limsup_{\varepsilon\to 0}{}^*(u - \phi^\varepsilon) = u - \phi$$

on $\hat{Q}$ we see that $(x_\varepsilon, t_\varepsilon) \to (L, \hat{t})$ as $\varepsilon \to 0$. Furthermore, since $h(\sigma) \to \infty$ as $\sigma \to 0$ we have that, for ε sufficiently small, $x_\varepsilon < L$ and $x_\varepsilon \in \Omega$. Because u is a subsolution for $F_{\min}$ this implies that

$$\phi_t^\varepsilon(x_\varepsilon, t_\varepsilon) \le F(x_\varepsilon, t_\varepsilon)\left(\phi_x^\varepsilon(x_\varepsilon, t_\varepsilon)^2 + \gamma^2\right)^{1/2},$$

which yields

$$\phi_t(x_\varepsilon, t_\varepsilon) \le F(x_\varepsilon, t_\varepsilon)\left(\left(\phi_x(x_\varepsilon, t_\varepsilon) - h'\left(\frac{d(x_\varepsilon)}{\varepsilon}\right)\right)^2 + \gamma^2\right)^{1/2}.$$

Observation 2. Set for $A > 0$, $\psi(x,t) := \phi(x,t) + Ad(x)$. Clearly

$$(u-\psi)(x,t) = (u-\phi)(x,t) - Ad(x) \le (u-\phi)(L,\hat{t}) = (u-\psi)(L,\hat{t})$$

for all $(x,t) \in \hat{Q}$. Since u is a subsolution for $F_{\min}$, we have

$$\min\left\{\tau - \hat{F}\hat{\alpha}, \tau - \hat{F}\left((p-A)^2 + \gamma^2\right)^{1/2}\right\} \le 0.$$

We treat the cases $\hat{F} > 0$ and $\hat{F} < 0$ separately.

Case I. $\hat{F} > 0$.

(i) $\gamma \geq \hat{\alpha}$. If $p \leq 0$, then $\tau \leq \hat{F}(p^2+\gamma^2)^{1/2} = \hat{F}\max\{\hat{\alpha}, (([p\nu]_-)^2+\gamma^2)^{1/2}\}$. If $p > 0$, then we use Observation 2, which yields, with $A = p$, $\tau \leq \hat{F}\gamma = \hat{F}\max\{\hat{\alpha}, (([p\nu]_-)^2+\gamma^2)^{1/2}\}$.

(ii) $\gamma < \hat{\alpha}$. If $p \leq -\delta$, then $\tau \leq \hat{F}(p^2+\gamma^2)^{1/2} = \hat{F}((p-\delta)(p+\delta)_- + \gamma^2+\delta^2)^{1/2}$. If $|p| < \delta$, then $\tau \leq \hat{F}\hat{\alpha} = \hat{F}\max\{\hat{\alpha}, (([p\nu]_-)^2+\gamma^2)^{1/2}\}$. If $p > \delta$, then we use Observation 2, which yields, with $A = p$, $\tau \leq \hat{F}\gamma \leq \hat{F}\max\{\hat{\alpha}, (([p\nu]_-)^2+\gamma^2)^{1/2}\}$.

Case II. $\hat{F} < 0$.

(i) $\gamma \geq \hat{\alpha}$ and $p \leq 0$. This is immediately treated by Observation 1. Sending ε to zero in

$$\phi_t(x_\varepsilon, t_\varepsilon) \leq F(x_\varepsilon, t_\varepsilon)(\phi_x^\varepsilon(x_\varepsilon, t_\varepsilon)^2+\gamma^2)^{1/2} \leq F(x_\varepsilon, t_\varepsilon)\gamma$$

yields $\tau \leq \hat{F}\gamma = \hat{F}\max\{\hat{\alpha}, (([-p\nu]_-)^2+\gamma^2)^{1/2}\}$.

(ii) $p\nu - \delta > 0$. Using Observation 1 we get that for small ε

$$\phi_x^\varepsilon(x_\varepsilon, t_\varepsilon) = \phi_x(x_\varepsilon, t_\varepsilon) - h'\left(\frac{d(x_\varepsilon)}{\varepsilon}\right) \geq \phi_x(x_\varepsilon, t_\varepsilon) > 0$$

so that

$$\phi_t(x_\varepsilon, t_\varepsilon) \leq F(x_\varepsilon, t_\varepsilon)\left(\phi_x^\varepsilon(x_\varepsilon, t_\varepsilon)^2+\gamma^2\right)^{1/2} \leq F(x_\varepsilon, t_\varepsilon)\left(\phi_x(x_\varepsilon, t_\varepsilon)^2+\gamma^2\right)^{1/2},$$

and sending $\varepsilon \to 0$,

$$\tau \leq \hat{F}(p^2+\gamma^2)^{1/2} = \hat{F}((p\nu+\delta)(p\nu-\delta)_+ + \gamma^2+\delta^2)^{1/2}.$$

(iii) $p\nu - \delta \leq 0$, $\gamma < \hat{\alpha}$. This is the remaining case. From Observation 2, by choosing A large we find

$$\tau \leq \hat{F}\hat{\alpha} = \hat{F}((p\nu+\delta)(p\nu-\delta)_+ + \gamma^2+\delta^2)^{1/2}. \qquad \square$$

Remark 1. As usual the definition of subsolutions (Definition 2.1) extends to an upper semicontinuous function u on $\hat{Q}$ provided that u is locally bounded on $\overline{Q}$. Similarly, a supersolution is defined for lower semicontinuous functions and not only for continuous functions. Results on equivalence (Lemmas 2.5, 2.6) are still valid for semicontinuous functions.

3. Comparison principle. Let Ω and T be as introduced in section 2 and set $\hat{Q} = \overline{\Omega} \times (0,T)$ and $\overline{Q} = \overline{\Omega} \times [0,T]$. In section 2 we defined a function G by

$$G(x,t,\tau,p) = \begin{cases} \tau - F(x,t)\left(p^2+\gamma^2\right)^{1/2} & \text{if} \quad x \in \Omega, \\ \tau - F(x,t)\max\left\{\alpha(x,t), \left(([p\nu(x)\mathrm{Sign}F(x,t)]_-)^2+\gamma^2\right)^{1/2}\right\} & \text{if} \quad x \in \partial\Omega, \end{cases}$$

where $\gamma \geq 0$ is a constant and $\nu(x)$ denotes the outer unit normal of $\partial\Omega$ (i.e., $\nu(0) = -1$, $\nu(L) = 1$).

THEOREM 3.1. *Assume that* $F \in C(\overline{Q})$, $\alpha \in C(\partial\Omega \times [0,T])$, *and*

$$|F(x,t) - F(y,t)| \leq C|x-y| \quad \textit{for all } (x,t),(y,t) \in \overline{Q} \tag{3.1}$$

holds for some constant $C > 0$ *independent of* t. *Let* u *and* $-v$ *be bounded upper semicontinuous functions on* $\overline{\Omega} \times [0,T)$. *Let* u *be a viscosity subsolution for* $G = 0$ *in*

$\hat{Q}$ and let v be a viscosity supersolution for $G = 0$ in $\hat{Q}$. If $u(\cdot, 0) \le v(\cdot, 0)$ in $\overline{\Omega}$, then $u \le v$ in $\hat{Q}$.

Proof. Suppose that the conclusion is false. Then there would exist a point $(x_1, t_1) \in \overline{\Omega} \times [0, T)$ such that $\mu := u(x_1, t_1) - v(x_1, t_1) > 0$. For positive parameters λ, β, δ, we set

$$\phi(x,t,y,s) = \lambda(x-y)^2 + \beta(t-s)^2 + \frac{\delta}{T-t} + \frac{\delta}{T-s},$$
$$\Phi(x,t,y,s) = u(x,t) - v(y,s) - \phi(x,t,y,s)$$

and, as preparation, study the behavior of a maximum point $(\hat{x}, \hat{t}, \hat{y}, \hat{s})$ of Φ on $\overline{Q} \times \overline{Q}$. We choose a small δ (fixed here) such that $0 < \delta < (T - t_1)\mu/4$. So we have

$$\max_{\overline{Q}\times\overline{Q}} \Phi \ge \mu/2 > 0, \tag{3.2}$$

and then $0 \le \hat{t}, \hat{s} < T$ holds uniformly for λ and β. Let $M > 0$ be an upper bound of both u and $-v$. By using (3.2) we see that

$$2M \ge u(\hat{x}, \hat{t}) - v(\hat{y}, \hat{s}) > \lambda|\hat{x} - \hat{y}|^2 + \beta|\hat{t} - \hat{s}|^2,$$

which leads to

$$\lambda|\hat{x} - \hat{y}|^2,\ \beta|\hat{t} - \hat{s}|^2 \text{ are bounded}, \tag{3.3}$$

and then $|\hat{x}-\hat{y}| \to 0$ (as $\lambda \to \infty$), $|\hat{t}-\hat{s}| \to 0$ (as $\beta \to \infty$), i.e., by taking a subsequence, there exists $(x_0, t_0) \in \hat{Q}$ such that

$$\hat{x}, \hat{y} \to x_0 \ (\text{as } \lambda \to \infty), \quad \hat{t}, \hat{s} \to t_0 \ (\text{as } \beta \to \infty). \tag{3.4}$$

(Because of our assumption $u \le v$ at $t = 0$, the time $t_0 > 0$ so that $\hat{t}$, $\hat{s} > 0$ for sufficiently large λ and β.) From now on, we use the same notation after taking a subsequence. By taking a subsequence, (3.3) implies that there are λ_0 and β_0 such that

$$\lambda|\hat{x} - \hat{y}|^2 \to \lambda_0 \ (\text{as } \lambda \to \infty), \quad \beta|\hat{t} - \hat{s}|^2 \to \beta_0 \ (\text{as } \beta \to \infty).$$

It follows that

$$\limsup_{\lambda\to\infty,\beta\to\infty} \max_{\overline{Q}\times\overline{Q}} \Phi \le u(x_0, t_0) - v(x_0, t_0) - \lambda_0 - \beta_0 - \frac{2\delta}{T - t_0}.$$

Since the left-hand side is equal to or greater than $u(x_0, t_0) - v(x_0, t_0) - 2\delta/(T - t_0)$, we have $\lambda_0 = \beta_0 = 0$, i.e.,

$$\lambda|\hat{x} - \hat{y}|^2 \to 0 \ (\text{as } \lambda \to \infty), \quad \beta|\hat{t} - \hat{s}|^2 \to 0 \ (\text{as } \beta \to \infty). \tag{3.5}$$

Now, let us start the main part of the proof. Note that our classifications given below are not disjoint but cover whole cases.

Case 1. $x_0 \in \Omega$. First, we discuss the case when $\hat{x}$, $\hat{y}$ converge to an interior point x_0 as $\lambda \to \infty$. We may assume that $\hat{x}$, $\hat{y} \in \Omega$. We use an abbreviated notation $\hat{\phi}_t = \phi_t(\hat{x}, \hat{t}, \hat{y}, \hat{s})$, etc. It follows from the definition of viscosity solutions that

$$\hat{\phi}_t - F(\hat{x}, \hat{t})(\hat{\phi}_x^2 + \gamma^2)^{1/2} \le 0 \quad \text{and} \quad (-\hat{\phi}_s) - F(\hat{y}, \hat{s})((-\hat{\phi}_y)^2 + \gamma^2)^{1/2} \ge 0.$$

Subtracting the second inequality from the first one we get

$$\begin{aligned}0 &\geq \hat{\phi}_t - F(\hat{x},\hat{t})(\hat{\phi}_x^2+\gamma^2)^{1/2} - \left((-\hat{\phi}_s) - F(\hat{y},\hat{s})((-\hat{\phi}_y)^2+\gamma^2)^{1/2}\right)\\ &\geq \frac{2\delta}{T^2} + (-F(\hat{x},\hat{t}) + F(\hat{y},\hat{s}))(4\lambda^2|\hat{x}-\hat{y}|^2+\gamma^2)^{1/2}.\end{aligned}$$

We send β to infinity and, after that, we send λ to infinity. Then, by using (3.5) and the Lipschitz continuity of $F(\cdot,t)$ as in (3.1), we see that the second term goes to zero. Since $\delta>0$, we get a contradiction.

Case 2. $x_0\in\partial\Omega$. Next, we discuss the case when $\hat{x}$, $\hat{y}$ go to a boundary point x_0 as $\lambda\to\infty$. If both $\hat{x}$ and $\hat{y}$ are in Ω for any large λ, we get a contradiction similar to that of Case 1. We classify the rest of Case 2 into three cases, Cases 2a, 2b, 2c, depending on the limit of the convergent subsequences. We further classify Cases 2b and 2c into more subcases as follows:

Case 2a	$\hat{x}=\hat{y}=x_0\in\partial\Omega$		
Case 2b	$\hat{x}\in\Omega\to x_0$, $\hat{y}=x_0\in\partial\Omega$	Case 2b(i)	$F(x_0,t_0)>0$
		Case 2b(ii)	$F(x_0,t_0)=0$
		Case 2b(iii)	$F(x_0,t_0)<0$
Case 2c	$\hat{x}=x_0\in\partial\Omega$, $\hat{y}\in\Omega\to x_0$	Case 2c(i)	$F(x_0,t_0)<0$
		Case 2c(ii)	$F(x_0,t_0)=0$
		Case 2c(iii)	$F(x_0,t_0)>0$

Since the proof for Case 2c and its subcases is symmetric to that of Case 2b, we do not present the proof for Case 2c.

Case 2a. When there exists a subsequence $\lambda\to\infty$ such that $\hat{x}=\hat{y}=x_0\in\partial\Omega$, it follows that

$$\begin{aligned}0 &\geq \hat{\phi}_t - F(\hat{x},\hat{t})\max\left\{\alpha(\hat{x},\hat{t}),\left(([\hat{\phi}_x\nu(\hat{x})\mathrm{Sign}F(\hat{x},\hat{t})]_-)^2+\gamma^2\right)^{1/2}\right\}\\ &\quad-\left((-\hat{\phi}_s) - F(\hat{y},\hat{s})\max\left\{\alpha(\hat{y},\hat{s}),\left(([(-\hat{\phi}_y)\nu(\hat{y})\mathrm{Sign}F(\hat{y},\hat{s})]_-)^2+\gamma^2\right)^{1/2}\right\}\right)\\ &\geq \frac{2\delta}{T^2} - F(x_0,\hat{t})\max\{\alpha(x_0,\hat{t}),\gamma\} + F(x_0,\hat{s})\max\{\alpha(x_0,\hat{s}),\gamma\},\end{aligned}$$

since $\hat{\phi}_x=-\hat{\phi}_y=0$. Since the second and third terms are continuous and since the sum of them goes to zero as $\beta\to\infty$ by (3.4), we get a contradiction.

Case 2b. When there exists a subsequence $\lambda\to\infty$ such that $\hat{x}\in\Omega$ and $\hat{y}=x_0\in\partial\Omega$, it follows that

(3.6)
$$\begin{aligned}&\hat{\phi}_t - F(\hat{x},\hat{t})(\hat{\phi}_x^2+\gamma^2)^{1/2}\leq 0,\\ &(-\hat{\phi}_s) - F(x_0,\hat{s})\max\left\{\alpha(x_0,\hat{s}),\left(([(-\hat{\phi}_y)\nu(x_0)\mathrm{Sign}F(x_0,\hat{s})]_-)^2+\gamma^2\right)^{1/2}\right\}\geq 0.\end{aligned}$$

We see that $(-\hat{\phi}_y)\nu(x_0)=2\lambda(\hat{x}-x_0)\nu(x_0)<0$. We shall classify this case into three subcases depending on the sign of F.

Case 2b(i). $F(x_0,t_0)>0$. When $F(x_0,t_0)>0$, we may assume that $F(x_0,\hat{s})>0$ holds for sufficiently large β. The second inequality of (3.6) implies that

$$\begin{aligned}0 &\leq (-\hat{\phi}_s) - F(x_0,\hat{s})\max\left\{\alpha(x_0,\hat{s}),(\hat{\phi}_y^2+\gamma^2)^{1/2}\right\}\\ &\leq (-\hat{\phi}_s) - F(x_0,\hat{s})(\hat{\phi}_y^2+\gamma^2)^{1/2}.\end{aligned}$$

We get a contradiction similar to that of Case 1.

Case 2b(ii). $F(x_0, t_0) = 0$. When $F(x_0, t_0) = 0$, we see that both $F(\hat{x}, \hat{t})$ and $F(x_0, \hat{s})$ go to zero as λ, $\beta \to \infty$. Then it is easy to get a contradiction from (3.6).

Case 2b(iii). $F(x_0, t_0) < 0$. When $F(x_0, t_0) < 0$, we may assume that $F(x_0, \hat{s}) < 0$ holds for any large β. The second inequality of (3.6) implies that

$$(-\hat{\phi}_s) - F(x_0, \hat{s}) \max\{\alpha(x_0, \hat{s}), \gamma\} \geq 0.$$

If $\alpha(x_0, t_0) \leq \gamma$, sending β to infinity, we have

$$0 \geq \frac{2\delta}{T^2} + (-F(\hat{x}, t_0) + F(x_0, t_0))(4\lambda^2|\hat{x} - x_0|^2 + \gamma^2)^{1/2}.$$

By using (3.5) and (3.1) the second term goes to zero as $\lambda \to \infty$. This yields a contradiction since $\delta > 0$.

Case R. We must discuss the remaining case when $\alpha_0 = \alpha(x_0, t_0) > \gamma$ from Case 2b(iii) (similarly from Case 2c(iii)). Let us change ϕ to a new test function ψ as follows:

$$\begin{aligned}\psi(x, t, y, s) &= \lambda \left(x - y + \mathrm{Sign} F(x_0, t_0) \frac{(\alpha_0^2 - \gamma^2)^{1/2} \nu(x_0)}{2\lambda} \right)^2 + \beta(t - s)^2 \\ &\quad + \frac{\delta}{T - t} + \frac{\delta}{T - s} \\ &= \lambda \left(x - y - \frac{(\alpha_0^2 - \gamma^2)^{1/2} \nu(x_0)}{2\lambda} \right)^2 + \beta(t - s)^2 + \frac{\delta}{T - t} + \frac{\delta}{T - s}.\end{aligned}$$

We then consider the maximum point $(\tilde{x}, \tilde{t}, \tilde{y}, \tilde{s})$ of

$$\Psi(x, t, y, s) = u(x, t) - v(y, s) - \psi(x, t, y, s)$$

on $\overline{Q} \times \overline{Q}$. It is easy to see that $0 < \tilde{t}, \tilde{s} < T$ holds uniformly for λ and β, $|\tilde{x} - \tilde{y}| \to 0$ (as $\lambda \to \infty$), and $|\tilde{t} - \tilde{s}| \to 0$ (as $\beta \to \infty$). We claim that

$$\tilde{x}, \tilde{y} \to x_0 \ (\text{as } \lambda \to \infty), \quad \tilde{t}, \tilde{s} \to t_0 \ (\text{as } \beta \to \infty). \tag{3.7}$$

If (3.7) does not hold, we see that, for some closed neighborhood K of (x_0, t_0), there exists $\mu' > 0$ such that

$$\Psi(\tilde{x}, \tilde{t}, \tilde{y}, \tilde{s}) - \max_{K \times K} \Psi \geq \mu'.$$

Since ψ is almost equal to ϕ as $\lambda \to \infty$ and $|x - y| \to 0$, we see that

$$\Phi(\tilde{x}, \tilde{t}, \tilde{y}, \tilde{s}) - \max_{K \times K} \Phi = \Phi(\tilde{x}, \tilde{t}, \tilde{y}, \tilde{s}) - \Phi(\hat{x}, \hat{t}, \hat{y}, \hat{s}) \geq \mu'/2$$

for large λ, which is inconsistent with the property that $(\hat{x}, \hat{t}, \hat{y}, \hat{s})$ is a global maximum point of Φ. Thus (3.7) holds.

We can discuss, similarly to Case 1 and Case 2a, respectively, the case when $\tilde{x}$, $\tilde{y} \in \Omega$ or $\tilde{x}$, $\tilde{y} \in \partial\Omega$ for large λ and obtain a contradiction. We thus study the remaining two cases, Cases Rb and Rc.

Case Rb. When there exists a subsequence $\lambda \to \infty$ such that $\tilde{x} \in \Omega$ and $\tilde{y} = x_0 \in \partial\Omega$, it follows that

(3.8)

$$\begin{aligned}&\tilde{\psi}_t - F(\tilde{x}, \tilde{t})(\tilde{\psi}_x^2 + \gamma^2)^{1/2} \leq 0, \\ &(-\tilde{\psi}_s) - F(x_0, \tilde{s}) \max\left\{ \alpha(x_0, \tilde{s}), \left(([(-\tilde{\psi}_y)\nu(x_0)\mathrm{Sign} F(x_0, \tilde{s})]_-)^2 + \gamma^2 \right)^{1/2} \right\} \geq 0.\end{aligned}$$

By the definition of ψ we see that $(-\tilde{\psi}_y)\nu(x_0) = 2\lambda(\tilde{x}-x_0)\nu(x_0) - (\alpha_0^2-\gamma^2)^{1/2} < 0$. In the remaining case from Case 2b(iii) we see that $F(x_0,t_0) < 0$, and then $F(x_0,\tilde{s}) < 0$ for large β. Thus the second inequality of (3.8) implies that

$$(-\tilde{\psi}_s) - F(x_0,\tilde{s})\max\{\alpha(x_0,\tilde{s}),\gamma\} \geq 0. \tag{3.9}$$

By the definition of ψ we also see that $(\gamma <)\ \alpha_0 \leq (\tilde{\psi}_x^2+\gamma^2)^{1/2}$. Subtracting (3.9) from the first inequality of (3.8) and sending $\beta \to \infty$, we have

$$0 \geq \frac{2\delta}{T^2} + (-F(\tilde{x},t_0) + F(x_0,t_0))\alpha_0$$

since $F(\tilde{x},t_0) < 0$ for large λ. By (3.7) and (3.1) the second term goes to zero as $\lambda \to \infty$. This yields a contradiction since $\delta > 0$.

Case Rc. When there exists a subsequence $\lambda \to \infty$ such that $\tilde{x} = x_0 \in \partial\Omega$ and $\tilde{y} \in \Omega$, it follows that

$$\tilde{\psi}_t - F(x_0,\tilde{t})\max\left\{\alpha(x_0,\tilde{t}), \left(([\tilde{\psi}_x\nu(x_0)\mathrm{Sign}F(x_0,\tilde{t})]_-)^2+\gamma^2\right)^{1/2}\right\} \leq 0,$$

$$(-\tilde{\psi}_s) - F(\tilde{y},\tilde{s})((-\tilde{\psi}_y)^2+\gamma^2)^{1/2} \geq 0.$$

If $\tilde{\psi}_x\nu(x_0) \geq 0$, the first inequality implies

$$\tilde{\psi}_t - F(x_0,\tilde{t})(\tilde{\psi}_x^2+\gamma^2)^{1/2} \leq 0$$

since $F(x_0,\tilde{t}) < 0$ for large β. We thus obtain a contradiction similar to that of Case 1. If $\tilde{\psi}_x\nu(x_0) < 0$, we have

$$0 \leq 2\lambda(x_0-\tilde{y})\nu(x_0) < (\alpha_0-\gamma^2)^{1/2},$$

and then $((-\tilde{\psi}_y)^2+\gamma^2)^{1/2} < \alpha_0$. Since $F(\tilde{y},\tilde{s}) < 0$ for large λ, β, we see that

$$0 \geq \frac{2\delta}{T^2} - F(x_0,\tilde{t})\alpha(x_0,\tilde{t}) + F(\tilde{y},\tilde{s})\alpha(x_0,t_0).$$

We get a contradiction by the continuity of the second and third terms. The proof of Theorem 3.1 is now complete. □

4. Existence theorem. Let Ω and T be introduced in section 2. Our goal is to show the existence of viscosity solutions of the dynamic boundary problem

$$\begin{cases} u_t - F\left(u_x^2+\gamma^2\right)^{1/2} = 0 & \text{in} \quad Q = \Omega\times(0,T), \\ u_t - F\alpha = 0 & \text{on} \quad \partial\Omega\times(0,T), \\ u|_{t=0} = a & \text{on} \quad \Omega. \end{cases} \tag{4.1}$$

THEOREM 4.1. *Assume that $F \in C^1(\overline{Q})$ and $\alpha \in C(\partial\Omega\times[0,T])$. Assume that a is a Lipschitz function over $\overline{\Omega}$. Then there exists a function $u \in C(\overline{Q})$ which is a unique viscosity solution of* (4.1). *Moreover, $|u_x|$ is bounded in $\overline{Q}$.*

Let $\varepsilon > 0$. First, we shall prove a priori estimates for a classical solution u^ε for the approximate problem

$$\begin{cases} u_t^\varepsilon - \varepsilon u_{xx}^\varepsilon = F^\varepsilon\left((u_x^\varepsilon)^2+\gamma^2\right)^{1/2} & \text{in} \quad Q, \\ u_t^\varepsilon + \varepsilon\nu u_x^\varepsilon = F^\varepsilon\max\left\{\alpha^\varepsilon, \left(([\nu u_x^\varepsilon\mathrm{Sign}F^\varepsilon]_-)^2+\gamma^2\right)^{1/2}\right\} & \text{on} \quad \partial\Omega\times(0,T), \\ u^\varepsilon|_{t=0} = a^\varepsilon & \text{on} \quad \Omega, \end{cases} \tag{4.2}$$

where ν denotes the outer unit normal of $\partial\Omega$. The existence of a solution of (4.2) shall be proved in the appendix (Theorem A.1).

PROPOSITION 4.2. *Assume that $F^\varepsilon \in C^1(\overline{Q}) \cap C^\infty(Q)$ and $\alpha^\varepsilon \in C^1(\partial\Omega \times [0,T])$. Assume that a^ε is a C^3 function over $\overline{\Omega}$ and $\varepsilon a^\varepsilon_{xx}$ is bounded on $\overline{\Omega}$ uniformly for ε. Let u^ε be a classical solution of* (4.2). *Then the estimate*

$$\max_{\overline{Q}} (|u^\varepsilon| + |u^\varepsilon_x| + |u^\varepsilon_t|) \le C \tag{4.3}$$

holds with some constant $C > 0$ depending only on T, γ, $|a^\varepsilon|_{C^1(\overline{\Omega})}$, $|\varepsilon a^\varepsilon_{xx}|_{C(\overline{\Omega})}$, $|F^\varepsilon|_{C^1(\overline{Q})}$, and $|\alpha^\varepsilon|_{C(\partial\Omega\times[0,T])}$.

Proof. We shall prove (4.3) by using maximum principles. In the proof we suppress superscripts of F^ε and α^ε to simplify the notation.

(i) *The estimate for u^ε.* We set $w(x,t) = e^{-t}u^\varepsilon(x,t)$. It follows from (4.2) that

$$\begin{cases} w_t + w - \varepsilon w_{xx} = e^{-t}F\left((e^t w_x)^2 + \gamma^2\right)^{1/2} & \text{in} \quad Q, \\ w_t + w + \varepsilon\nu w_x \\ \qquad = e^{-t}F\max\left\{\alpha, \left(([e^t\nu w_x \mathrm{Sign} F]_-)^2 + \gamma^2\right)^{1/2}\right\} & \text{on} \quad \partial\Omega\times(0,T), \\ w(x,0) = a^\varepsilon(x) & \text{for} \quad x \in \Omega. \end{cases} \tag{4.4}$$

Assume that w has a positive maximum in $\overline{Q}$. Let $(\hat{x},\hat{t})$ be the maximum point of w and let λ be its maximum value, i.e., $\max_{\overline{Q}} w = w(\hat{x},\hat{t})$ and $\lambda = w(\hat{x},\hat{t})$. We assume that

$$\lambda > \max_{\overline{\Omega}} |a^\varepsilon| \quad \text{and} \quad \lambda > \max_{\overline{Q}} |F| \max\left\{\max_{\partial\Omega\times[0,T]} |\alpha|, \gamma\right\} \tag{4.5}$$

and shall show that it is inconsistent with (4.4) for w. First, we observe that $\hat{t} > 0$ by the first inequality of (4.5) and obtain $w_t(\hat{x},\hat{t}) \ge 0$. When $\hat{x} \in \partial\Omega$, we also observe that $\nu(\hat{x})w_x(\hat{x},\hat{t}) \ge 0$. If $F(\hat{x},\hat{t}) > 0$, then it follows from the second identity of (4.4) that

$$\lambda \le e^{-\hat{t}}F(\hat{x},\hat{t})\max\{\alpha(\hat{x},\hat{t}),\gamma\},$$

which is a contradiction of the second inequality of (4.5). On the other hand, $F(\hat{x},\hat{t}) \le 0$ implies that $\lambda \le 0$, but it is also a contradiction. When $\hat{x} \in \Omega$, we see that $w_x(\hat{x},\hat{t}) = 0$, $w_{xx}(\hat{x},\hat{t}) \le 0$ to get $\lambda \le e^{-\hat{t}}F(\hat{x},\hat{t})\gamma$ by the first identity of (4.4), which is a contradiction of the second inequality of (4.5). Thus we have an a priori estimate,

$$\max_{\overline{Q}} u^\varepsilon \le e^T \max\left\{\max_{\overline{\Omega}} |a^\varepsilon|, \max_{\overline{Q}} |F| \max\left\{\max_{\partial\Omega\times[0,T]} |\alpha|, \gamma\right\}\right\}.$$

We argue in the same way for a negative minimum to get

$$\max_{\overline{Q}} |u^\varepsilon| \le e^T \max\left\{\max_{\overline{\Omega}} |a^\varepsilon|, \max_{\overline{Q}} |F| \max\left\{\max_{\partial\Omega\times[0,T]} |\alpha|, \gamma\right\}\right\}. \tag{4.6}$$

(ii) *The estimate for u^ε_x.* We set $w(x,t) = e^{-Kt}u^\varepsilon_x$ and $K = 2\max_{\overline{Q}} |F_x|$. It

follows from (4.2) that

(4.7)
$$\begin{cases} w_t + Kw - \varepsilon w_{xx} \\ \qquad = e^{-Kt}F_x\left((e^{Kt}w)^2+\gamma^2\right)^{1/2} + \dfrac{Fe^{Kt}ww_x}{((e^{Kt}w)^2+\gamma^2)^{1/2}} & \text{in} \quad Q, \\ \varepsilon w_x + e^{-Kt}F\left((e^{Kt}w)^2+\gamma^2\right)^{1/2} + \varepsilon\nu w \\ \qquad = e^{-Kt}F\max\left\{\alpha, \left(([e^{Kt}\nu w \mathrm{Sign} F]_-)^2+\gamma^2\right)^{1/2}\right\} & \text{on} \quad \partial\Omega\times(0,T), \\ w(x,0) = a_x^\varepsilon(x) & \text{for} \quad x\in\Omega. \end{cases}$$

As before we take $(\hat{x},\hat{t})$, λ satisfying $\max_{\overline{Q}} w = w(\hat{x},\hat{t})$ and $\lambda = w(\hat{x},\hat{t})$. We assume that

$$\lambda > \max_{\overline{\Omega}} |a_x^\varepsilon|, \quad \lambda > \max_{\partial\Omega\times[0,T]} |\alpha|, \quad \text{and} \quad \lambda > \gamma \tag{4.8}$$

and shall show that it is inconsistent with (4.7) for w. First, we see that $\hat{t} > 0$ by the first inequality of (4.8) and observe $w_t(\hat{x},\hat{t}) \geq 0$. When $\hat{x}\in\partial\Omega$, we also see that $\nu(\hat{x})w_x(\hat{x},\hat{t}) \geq 0$. If $\nu(\hat{x})F(\hat{x},\hat{t}) > 0$, then it follows from the second identity of (4.7) after multiplying $\nu(\hat{x})$ on both sides that

$$\nu(\hat{x})e^{-K\hat{t}}F(\hat{x},\hat{t})\left((e^{K\hat{t}}\lambda)^2+\gamma^2\right)^{1/2} + \varepsilon\lambda \leq \nu(\hat{x})e^{-K\hat{t}}F(\hat{x},\hat{t})\max\{\alpha(\hat{x},\hat{t}),\gamma\}.$$

Dividing both sides by $\nu(\hat{x})e^{-K\hat{t}}F(\hat{x},\hat{t}) > 0$, we obtain

$$\left((e^{K\hat{t}}\lambda)^2+\gamma^2\right)^{1/2} + \frac{\varepsilon\lambda}{\nu(\hat{x})e^{-K\hat{t}}F(\hat{x},\hat{t})} \leq \max\{\alpha(\hat{x},\hat{t}),\gamma\}.$$

The left-hand side is strictly greater than λ, which is a contradiction of the second and third inequalities of (4.8). If $\nu(\hat{x})F(\hat{x},\hat{t}) < 0$, then a similar calculation shows that

$$\begin{aligned} &\nu(\hat{x})e^{-K\hat{t}}F(\hat{x},\hat{t})\left((e^{K\hat{t}}\lambda)^2+\gamma^2\right)^{1/2} + \varepsilon\lambda \\ &\qquad \leq \nu(\hat{x})e^{-K\hat{t}}F(\hat{x},\hat{t})\max\left\{\alpha(\hat{x},\hat{t}), ((e^{K\hat{t}}\lambda)^2+\gamma^2)^{1/2}\right\} \end{aligned}$$

and, since $\nu(\hat{x})e^{-K\hat{t}}F(\hat{x},\hat{t}) < 0$,

$$\left((e^{K\hat{t}}\lambda)^2+\gamma^2\right)^{1/2} + \frac{\varepsilon\lambda}{\nu(\hat{x})e^{-K\hat{t}}F(\hat{x},\hat{t})} \geq \max\left\{\alpha(\hat{x},\hat{t}), ((e^{K\hat{t}}\lambda)^2+\gamma^2)^{1/2}\right\}.$$

Since the second term of the left-hand side is negative, we get a contradiction. The term $\varepsilon\nu u_x$ plays a crucial role in getting a contradiction for the case $\nu(\hat{x})F(\hat{x},\hat{t}) \neq 0$. If $F(\hat{x},\hat{t}) = 0$, then the second identity of (4.7) implies $\lambda \leq 0$, which is also a contradiction. When $\hat{x}\in\Omega$, we see that $w_x(\hat{x},\hat{t}) = 0$, $w_{xx}(\hat{x},\hat{t}) \leq 0$ and then

$$K\lambda \leq e^{-K\hat{t}}F_x(\hat{x},\hat{t})\left((e^{K\hat{t}}\lambda)^2+\gamma^2\right)^{1/2}$$

from the first identity of (4.7). The right-hand side is less than or equal to $K(\lambda^2+\gamma^2)^{1/2}/2$, which is a contradiction of the third inequality of (4.8). Hence, we have the estimate

$$\max_{\overline{Q}} |u_x^\varepsilon| \leq e^{KT}\max\left\{\max_{\overline{\Omega}} |a_x^\varepsilon|, \max_{\partial\Omega\times[0,T]} |\alpha|, \gamma\right\}, \quad K = 2\max_{\overline{Q}} |F_x|. \tag{4.9}$$

(iii) *The estimate for u_t^ε.* We set $w(x,t) = e^{-\mu t} u_t^\varepsilon(x,t)$ ($\mu = 1$). It follows from (4.2) that

$$
\begin{cases}
w_t + \mu w - \varepsilon w_{xx} & \\
\qquad = e^{-\mu t} F_t \left((u_x^\varepsilon)^2 + \gamma^2 \right)^{1/2} + \dfrac{F u_x^\varepsilon w_x}{((u_x^\varepsilon)^2 + \gamma^2)^{1/2}} & \text{in} \quad Q, \\
w + \varepsilon e^{-\mu t} \nu u_x^\varepsilon & \\
\qquad = e^{-\mu t} F \max \left\{ \alpha, \left(([\nu u_x^\varepsilon \mathrm{Sign} F]_-)^2 + \gamma^2 \right)^{1/2} \right\} & \text{on} \quad \partial\Omega \times (0,T), \\
w(x,0) = \varepsilon a_{xx}^\varepsilon(x) + F(x,0) \left(a_x^\varepsilon(x)^2 + \gamma^2 \right)^{1/2} & \text{for} \quad x \in \Omega.
\end{cases} \tag{4.10}
$$

Let $(\hat{x}, \hat{t})$ and λ satisfy $\max_{\overline{Q}} w = w(\hat{x}, \hat{t})$ and $\lambda = w(\hat{x}, \hat{t})$. We assume that

$$
\lambda > \Gamma_0, \quad \lambda > \Gamma_1, \quad \text{and} \quad \lambda > \Gamma_2, \tag{4.11}
$$

with

$$
\begin{aligned}
\Gamma_0 &= \varepsilon \max_{\overline{\Omega}} |a_{xx}^\varepsilon| + \max_{\overline{Q}} |F| \left(\max_{\overline{\Omega}} |a_x^\varepsilon| + \gamma \right), \\
\Gamma_1 &= \max_{\overline{Q}} |F| \max \left\{ \max_{\partial\Omega \times [0,T]} |\alpha|, \max_{\overline{Q}} |u_x^\varepsilon| + \gamma \right\}, \\
\Gamma_2 &= \max_{\overline{Q}} |F_t| \left(\max_{\overline{Q}} |u_x^\varepsilon| + \gamma \right).
\end{aligned}
$$

We shall show that it is inconsistent with (4.10) for w. First, we see that $\hat{t} > 0$ by the first inequality of (4.11) and obtain $w_t(\hat{x}, \hat{t}) \geq 0$. When $\hat{x} \in \partial\Omega$, by the second identity of (4.10) we see that

$$
\lambda + \varepsilon e^{-\mu \hat{t}} \nu(\hat{x}) u_x^\varepsilon(\hat{x}, \hat{t}) \leq e^{-\mu \hat{t}} F(\hat{x}, \hat{t}) \max \left\{ \alpha(\hat{x}, \hat{t}), (u_x^\varepsilon(\hat{x}, \hat{t})^2 + \gamma^2)^{1/2} \right\}.
$$

Since u_x^ε is bounded uniformly for ε, by choosing ε sufficiently small at the second term of the left-hand side, we get a contradiction to the second inequality of (4.11). When $\hat{x} \in \Omega$, we see that $w_x(\hat{x}, \hat{t}) = 0$, $w_{xx}(\hat{x}, \hat{t}) \leq 0$ and then

$$
\mu \lambda \leq e^{-\mu \hat{t}} F_t(\hat{x}, \hat{t}) ((u_x^\varepsilon)^2 + \gamma^2)^{1/2}
$$

from the first identity of (4.10), which is a contradiction of the third inequality of (4.11). Hence, we have the estimate

$$
\max_{\overline{Q}} |u_t^\varepsilon| \leq e^T \max \{ \Gamma_0, \Gamma_1, \Gamma_2 \}. \tag{4.12}
$$

By (4.6), (4.9), (4.12) we get the estimate (4.3) and now complete the proof of Proposition 4.2. □

Remark 2. (1) When $F_x^\varepsilon \equiv 0$, one can choose any $K > 0$ in the proof (ii). So, we have

$$
\max_{\overline{Q}} |u_x^\varepsilon| \leq \max \left\{ \max_{\overline{\Omega}} |a_x^\varepsilon|, \max_{\partial\Omega \times [0,T]} |\alpha|, \gamma \right\}. \tag{4.9$'$}
$$

(2) When $F_t^\varepsilon \equiv 0$, we can choose any $\mu > 0$ in the proof (iii). So, we have

$$
\max_{\overline{Q}} |u_t^\varepsilon| \leq \max \{ \Gamma_0, \Gamma_1, \Gamma_2 \}. \tag{4.12$'$}
$$

(3) To carry out the above proof we implicitly invoke the regularity u^ε, u^ε_x, u^ε_{xx}, $u^\varepsilon_t \in C(\overline{Q})$ together with u^ε_{xt}, u^ε_{xxx}, u^ε_{tt}, $u^\varepsilon_{xxt} \in C(Q)$. In Proposition 4.2 and Corollary 4.3 as a classical solution we require at least this regularity.

In a way similar to the proof of Proposition 4.2 one is able to prove an a priori estimate, which is useful in proving the global existence of solutions of (4.2) (cf. Theorem A.1).

COROLLARY 4.3. *Assume that F^ε and α^ε are in Proposition* 4.2 *and that a is a C^3 function over $\overline{\Omega}$. Let $\sigma \in [0,1]$ and $\varepsilon > 0$. Let v be a classical solution of*

$$\begin{cases} v_t - \varepsilon v_{xx} = \sigma F^\varepsilon \left((v_x)^2 + \gamma^2\right)^{1/2} & \text{in} \quad Q, \\ v_t = \sigma F^\varepsilon \max\left\{\alpha^\varepsilon, \left(([\nu v_x \mathrm{Sign} F^\varepsilon]_-)^2 + \gamma^2\right)^{1/2}\right\} - \sigma\varepsilon\nu v_x & \text{on} \quad \partial\Omega \times (0,T), \\ v|_{t=0} = \sigma a & \text{on} \quad \Omega. \end{cases}$$

Then the estimate

$$\max_{\overline{Q}} (|v| + |v_x|) \le C \tag{4.13}$$

holds with some constant $C > 0$ independent of $\varepsilon \in (0,1)$ and $\sigma \in [0,1]$.

Proof of Theorem 4.1. For a given Lipschitz function a there is a sequence $a^\varepsilon \in C^\infty(\overline{\Omega})$ such that $a^\varepsilon \to a$ uniformly and that $|a^\varepsilon_x|_{C(\overline{\Omega})}$ and $|\varepsilon a^\varepsilon_{xx}|_{C(\overline{\Omega})}$ are bounded. For a given $F \in C^1(\overline{Q})$ and $\alpha \in C(\partial\Omega \times [0,T])$ there is a sequence $\{F^\varepsilon, \alpha^\varepsilon\}$ with $F^\varepsilon \in C^1(\overline{Q}) \cap C^\infty(Q)$, $\alpha^\varepsilon \in C^1(\partial\Omega \times [0,T])$ such that $F^\varepsilon \to F$ uniformly in $\overline{Q}$ and $\alpha^\varepsilon \to \alpha$ uniformly in $\partial\Omega \times [0,T]$ and that $|F^\varepsilon|_{C^1(\overline{Q})}$ and $|\alpha^\varepsilon|_{C(\partial\Omega\times[0,T])}$ are bounded as $\varepsilon \to 0$. By Theorem A.1 there exists a unique classical solution u^ε of (4.2).

By the uniform estimate (4.3) the Arzelà–Ascoli theorem implies that there exists a function u such that

$$u^\varepsilon \to u \quad \text{uniformly on } \overline{Q}.$$

We shall show that u is the viscosity solution of the original dynamic boundary problem (4.1). Since the proof for viscosity supersolutions is symmetric, we only prove that u is a viscosity subsolution for $G = 0$. To do this, let $\phi \in C^2(\overline{Q})$ be a test function and let $(\hat{x}, \hat{t}) \in \hat{Q} = \overline{\Omega} \times (0,T)$ be the maximum point of $u - \phi$. We may assume that $(\hat{x}, \hat{t})$ is a strict maximum of $u - \phi$. Then there exists $(x_\varepsilon, t_\varepsilon)$ such that $(x_\varepsilon, t_\varepsilon) \to (\hat{x}, \hat{t})$ and $\sup_{\hat{Q}}(u^\varepsilon - \phi) = (u^\varepsilon - \phi)(x_\varepsilon, t_\varepsilon)$.

Case 1. When there exists a subsequence $\{(x_\varepsilon, t_\varepsilon) \in Q\}$, we see that $u^\varepsilon_t(x_\varepsilon, t_\varepsilon) = \phi_t(x_\varepsilon, t_\varepsilon)$, $u^\varepsilon_x(x_\varepsilon, t_\varepsilon) = \phi_x(x_\varepsilon, t_\varepsilon)$, and $u^\varepsilon_{xx}(x_\varepsilon, t_\varepsilon) \le \phi_{xx}(x_\varepsilon, t_\varepsilon)$. Since u^ε satisfies the first identity of (4.2) at $(x_\varepsilon, t_\varepsilon)$ as a classical solution, we get

$$\phi_t(x_\varepsilon, t_\varepsilon) - \varepsilon\phi_{xx}(x_\varepsilon, t_\varepsilon) - F^\varepsilon(x_\varepsilon, t_\varepsilon)(\phi_x(x_\varepsilon, t_\varepsilon)^2 + \gamma^2)^{1/2} \le 0.$$

By $\varepsilon \to 0$ we see that u is a viscosity subsolution at $(\hat{x}, \hat{t})$.

Case 2. When $\hat{x} \in \partial\Omega$ and there is a subsequence $\{(x_\varepsilon, t_\varepsilon) \in \partial\Omega \times (0,T)\}$, we see that $u^\varepsilon_t(x_\varepsilon, t_\varepsilon) = \phi_t(x_\varepsilon, t_\varepsilon)$ and $\nu u^\varepsilon_x(x_\varepsilon, t_\varepsilon) \ge \nu\phi_x(x_\varepsilon, t_\varepsilon)$. Since u^ε satisfies the second identity of (4.2) at $(x_\varepsilon, t_\varepsilon)$ as a classical solution, we get

$$\begin{aligned} &\phi_t(x_\varepsilon, t_\varepsilon) + \varepsilon\nu\phi_x(x_\varepsilon, t_\varepsilon) \\ &\quad \le F^\varepsilon(x_\varepsilon, t_\varepsilon) \max\left\{\alpha^\varepsilon(x_\varepsilon, t_\varepsilon), \left(([\nu u^\varepsilon_x(x_\varepsilon, t_\varepsilon)\mathrm{Sign} F^\varepsilon(x_\varepsilon, t_\varepsilon)]_-)^2 + \gamma^2\right)^{1/2}\right\}. \end{aligned} \tag{4.14}$$

If $F(\hat{x},\hat{t}) > 0$, we may assume that $F^\varepsilon(x_\varepsilon,t_\varepsilon) > 0$. We also see that $([\nu u_x^\varepsilon(x_\varepsilon,t_\varepsilon)]_-)^2 \le \phi_x(x_\varepsilon,t_\varepsilon)^2$ and then

$$\phi_t(x_\varepsilon,t_\varepsilon) + \varepsilon\nu\phi_x(x_\varepsilon,t_\varepsilon) \le F^\varepsilon(x_\varepsilon,t_\varepsilon)\max\left\{\alpha^\varepsilon(x_\varepsilon,t_\varepsilon), \left(\phi_x(x_\varepsilon,t_\varepsilon)^2+\gamma^2\right)^{1/2}\right\}.$$

By $\varepsilon \to 0$ it holds that u satisfies either the first or second identity of (4.1) as a viscosity subsolution at $(\hat{x},\hat{t})$. If $F(\hat{x},\hat{t}) < 0$, we may assume that $F^\varepsilon(x_\varepsilon,t_\varepsilon) < 0$. It is easy to see that

$$\phi_t(x_\varepsilon,t_\varepsilon) + \varepsilon\nu\phi_x(x_\varepsilon,t_\varepsilon) \le F^\varepsilon(x_\varepsilon,t_\varepsilon)\alpha^\varepsilon(x_\varepsilon,t_\varepsilon).$$

By $\varepsilon \to 0$ we get the second identity of (4.1) as a viscosity subsolution at $(\hat{x},\hat{t})$. If $F(\hat{x},\hat{t}) = 0$, the right-hand side of (4.14) vanishes as $\varepsilon \to 0$. We get $u_t \le 0$ in the viscosity sense at $(\hat{x},\hat{t})$. Thus u is a viscosity solution of (4.1), and it is unique by the comparison principle (Theorem 3.1).

The Lipschitz continuity of u in x follows from (4.9). □

5. Relation to other boundary conditions. We shall relate an inhomogeneous Neumann boundary value problem for

$$u_t - F(u_x^2+\gamma^2)^{1/2} = 0 \tag{5.1}$$

supplemented with the dynamic boundary

$$u_t - F\alpha = 0. \tag{5.2}$$

Formally, (5.1) and (5.2) yields

$$F(u_x^2+\gamma^2)^{1/2} = F\alpha.$$

If F is not zero, this implies $u_x^2+\gamma^2 = \alpha^2$. Thus we obtain

$$\partial u/\partial\nu = u_x\nu = \pm(\alpha^2-\gamma^2)^{1/2} \tag{5.3}$$

on the boundary. The Neumann data in (5.3) needs more explanation since both its sign and its value for $\alpha^2 < \gamma^2$ are unclear. We shall clarify these points and prove that a solution of (5.1), (5.2) solves an inhomogeneous Neumann problem in the viscosity sense (Theorem 5.1).

When we are asked to solve (5.1) and (5.2), we are tempted to integrate (5.2) in order to obtain the Dirichlet condition:

$$u(x,t) = \int_0^t F(x,\tau)\alpha(x,\tau)d\tau + a(x), \quad x \in \partial\Omega. \tag{5.4}$$

However, (5.1) with the Dirichlet condition (5.4) is not, unfortunately, equivalent to (5.1), (5.2). We shall give a counterexample in the last part of this section.

THEOREM 5.1. *Assume that F and α are continuous on $\overline{Q}$ and $\partial\Omega\times[0,T]$, respectively. Assume that u is a viscosity subsolution (resp., supersolution) for G in $\hat{Q}$. Then u is a viscosity subsolution (resp., supersolution) of the Neumann problem of* (5.1) *in $\hat{Q}$ with*

$$\partial u/\partial\nu = \mathrm{Sign}F\left\{(\alpha-\gamma)_+(\alpha+\gamma)\right\}^{1/2}. \tag{5.5}$$

Here β_+ is the plus part of β defined by $\beta_+ = \max(\beta, 0)$.

Proof. We suppress the word viscosity in the proof. Since the proof for supersolutions is symmetric, we shall present the proof for subsolutions only. We may assume that u is upper semicontinuous in $\hat{Q}$. Assume that u is a subsolution for G in $\hat{Q}$. Assume that $u - \phi$ takes its maximum over $\hat{Q}$ at $(\hat{x}, \hat{t})$ with $\hat{x} \in \overline{\Omega}$, $\hat{t} \in (0, T)$ for $\phi \in C^1(\hat{Q})$. We may assume that $\hat{x} \in \partial\Omega$ since the equation is the same in $\Omega \times (0, T)$. To simplify notation we set

$$\tau = \phi_t(\hat{x}, \hat{t}), \quad p = \phi_x(\hat{x}, \hat{t}), \quad \hat{F} = F(\hat{x}, \hat{t}), \quad \hat{\alpha} = \alpha(\hat{x}, \hat{t}).$$

We have to prove that

$$\min\left\{\tau - \hat{F}(p^2 + \gamma^2)^{1/2}, p\nu - \mathrm{Sign}\hat{F}\left\{(\hat{\alpha} - \gamma)_+(\hat{\alpha} + \gamma)\right\}^{1/2}\right\} \leq 0. \tag{5.6}$$

To prove (5.6) we may assume that

$$\tau - \hat{F}(p^2 + \gamma^2)^{1/2} > 0. \tag{5.7}$$

Case 1 ($\hat{F} < 0$). Since u is a subsolution of $G = 0$, we have

$$\tau - \hat{F}\max\left\{\hat{\alpha}, \left(([p\nu]_+)^2 + \gamma^2\right)^{1/2}\right\} \leq 0. \tag{5.8}$$

From (5.7) and (5.8) it follows that

$$\hat{F}\left(p^2 + \gamma^2\right)^{1/2} < \hat{F}\max\left\{\hat{\alpha}, \left(([p\nu]_+)^2 + \gamma^2\right)^{1/2}\right\}$$

or

$$\begin{aligned}\left(p^2 + \gamma^2\right)^{1/2} &> \max\left\{\hat{\alpha}, \left(([p\nu]_+)^2 + \gamma^2\right)^{1/2}\right\} \\ &\geq \left(([p\nu]_+)^2 + \gamma^2\right)^{1/2}.\end{aligned} \tag{5.9}$$

This implies $p^2 > ([p\nu]_+)^2$, so we obtain

$$p\nu < 0. \tag{5.10}$$

Assume that $\hat{\alpha} > \gamma$. From (5.9) it follows that $(p^2 + \gamma^2)^{1/2} > \hat{\alpha}$. This together with (5.10) implies that

$$p\nu < -(\hat{\alpha}^2 - \gamma^2)^{1/2} \quad \text{for } \hat{\alpha} > \gamma. \tag{5.11}$$

By (5.10) and (5.11) we obtain

$$p\nu < -\left\{(\hat{\alpha} - \gamma)_+(\hat{\alpha} + \gamma)\right\}^{1/2}.$$

We now obtain (5.6) when $\hat{F} < 0$.

Case 2 ($\hat{F} > 0$). We note that $G = 0$ is equivalent to the dynamic boundary value problem (5.1), (5.2). Since (5.7) holds, we have

$$\tau - \hat{F}\hat{\alpha} \leq 0. \tag{5.12}$$

From (5.7) and (5.12) it follows that

$$\hat{F}(p^2 + \gamma^2)^{1/2} < \hat{F}\hat{\alpha}.$$

Since $\hat{F} > 0$, this yields $(p^2 + \gamma^2)^{1/2} < \hat{\alpha}$ and implies

$$p^2 < \hat{\alpha}^2 - \gamma^2 \quad \text{or} \quad |p|^2 \le \hat{\alpha}^2 - \gamma^2 = (\hat{\alpha} - \gamma)_+(\hat{\alpha} + \gamma).$$

We have thus proved (5.6) when $\hat{F} > 0$.

Case 3 ($\hat{F} = 0$). Since $G = 0$ is equivalent to the dynamic boundary value problem (5.1), (5.2), $\tau \le 0$ is always fulfilled if $\hat{F} = 0$. Thus we have proved (5.6). □

We shall give a counterexample to show that the problem (5.1), (5.2) is different from the Dirichlet problem (5.1), (5.4) in the viscosity sense. We suppress the word viscosity.

We shall give two different functions u and v which initially agree with each other, but u solves (5.1), (5.2) while v solves (5.1), (5.4) when $\alpha \equiv 1$, $F \equiv 1$, $\alpha > \gamma$, and $\Omega = (0, \infty)$. Although it is not difficult to give such functions for $\Omega = (0, L)$ with more general α and F, we keep such assumptions to clarify the argument. Let β be a constant strictly greater than $\sigma = (1 - \gamma^2)^{1/2}$ so that $\eta = (\beta^2 + \gamma^2)^{1/2} > 1$. We set

$$w(x,t) = \min\{\beta + \gamma t, \beta x + \eta t, -\sigma x + \sigma + \beta + t\}, \quad x \in \overline{\Omega}. \tag{5.13}$$

This function is nondecreasing in t and

$$w(x,0) = \min\{\beta x, -\sigma x + \sigma + \beta\}$$

so that $w(x,0)$ is linear except at $x = 1$. At time $t_0 = \beta(\eta - \gamma)^{-1}$

$$w(x,t_0) = \min\{\beta + \gamma t_0, -\sigma x + \sigma + \beta + t_0\}.$$

Since $\beta \ge \sigma$, it is easy to see that

$$\phi_t - \left(\phi_x^2 + \gamma^2\right)^{1/2} \le 0 \quad \text{at } (\hat{x}, \hat{t})$$

if $w - \phi$ attains its maximum at $(\hat{x}, \hat{t})$ over $\overline{\Omega} \times (0, t_0]$ even if $\hat{x} \in \partial\Omega$. So w is a subsolution of $\overline{\Omega} \times (0, t_0]$ of (5.1), (5.2) and (5.1), (5.4). It is easy to see that w is a supersolution of (5.1), (5.2) and (5.1), (5.4) in $\overline{\Omega} \times (0, t_0]$ since $w_t \ge 1$, $w \ge t_0$ on the boundary. We now set

$$u(x,t) = v(x,t) = w(x,t) \quad \text{for } t \le t_0, x \in \overline{\Omega} \tag{5.14}$$

and

$$v(x,t) = \min\{\beta + \gamma t, -\sigma x + \sigma + \beta + t\} \quad \text{for } t \ge t_0, x \in \overline{\Omega}, \tag{5.15}$$

$$u(x,t) = \max\{\beta + (\gamma - 1)t_0 + t - \sigma x, v(x,t)\} \quad \text{for } t \ge t_0, x \in \overline{\Omega}. \tag{5.16}$$

As for w it is easy to see that v is a subsolution of both the dynamic (5.1), (5.2) and the Dirichlet problem (5.1), (5.4) in $\overline{\Omega} \times (0, \infty)$. Since $\eta > 1$ so that $t_1 = \beta(1 - \gamma)^{-1} > t_0$, and since $v(0,t) > t$ for $t < t_1$, v is a supersolution of the Dirichlet problem in $\overline{\Omega} \times (0, t_1)$. However, v is not a supersolution in $\overline{\Omega} \times (0, t_1)$ of (5.1), (5.2) since at the boundary $v_t < 1$ with $v_x = 0$.

Since $u_t = 1$ on the boundary and since it is easy to see that u is a solution of (5.1) in $\Omega \times (0, \infty)$, we conclude that u is a solution of (5.1), (5.2) in $\overline{\Omega} \times (0, \infty)$. This is not a subsolution of (5.1), (5.4) in $\overline{\Omega} \times (0, \infty)$ since $u(0,t) > t$ by $\eta > 1$ and

$$\phi_t - \left(\phi_x^2 + \gamma^2\right)^{1/2} > 0 \quad \text{at } (0, \hat{t})$$

if $u-\phi$ attains its maximum on $\overline{\Omega}\times(0,\infty)$ and $\hat{t}>t_0$. (The function u is a supersolution of (5.1), (5.4) since $u(0,t)>t$.) We summarize our results.

PROPOSITION 5.2. *Assume that $\alpha\equiv F\equiv 1$ and $\gamma<1$. Let $\beta>\sigma=(1-\gamma^2)^{1/2}$. For $\Omega=(0,\infty)$, let u and v be functions defined by* (5.13)–(5.16). *Then u is a solution of the dynamic boundary problem* (5.1), (5.2) *in $\overline{\Omega}\times(0,\infty)$ while v is a solution of the Dirichlet problem* (5.1), (5.4) *in $\overline{\Omega}\times(0,t_1)$ with $t_1=\beta(1-\gamma)^{-1}$. However, u is not a subsolution of* (5.1), (5.4) *in $\overline{\Omega}\times(0,T)$, $T>t_0$, while u is a supersolution of* (5.1), (5.4) *in $\overline{\Omega}\times(0,\infty)$. The function v is not a supersolution of* (5.1), (5.2) *while it is a subsolution of* (5.1), (5.2).

Appendix. Existence of solutions of approximate solutions. Our goal is to prove the following theorem.

THEOREM A.1. *For $T>0$ assume that $F\in C^1(\overline{Q})\cap C^\infty(Q)$ and $\alpha\in C^1(\partial\Omega\times[0,T])$ with $Q=\Omega\times(0,T)$, where Ω is a bounded open interval. Assume that $a\in C^3(\overline{\Omega})$ and $\gamma\in\mathbb{R}$. Then for each $\varepsilon>0$ there exists a solution $u\in C^{2,1}(\overline{Q})\cap C^\infty(Q)$ of*

$$\begin{cases} u_t-\varepsilon u_{xx}=F(u_x^2+\gamma^2)^{1/2} & \text{in} \quad Q,\\ u_t+\varepsilon\nu u_x=F\max\left\{\alpha,\left(([\nu u_x\mathrm{Sign}F]_-)^2+\gamma^2\right)^{1/2}\right\} & \text{on} \quad \partial\Omega\times(0,T),\\ u|_{t=0}=a & \text{on} \quad \Omega. \end{cases}$$

The space $C^{2,1}(\overline{Q})$ denotes the space of all $u\in C(\overline{Q})$ satisfying u_x, u_{xx}, $u_t\in C(\overline{Q})$. The space $X=C^{1,0}(\overline{Q})$ denotes the space of all $u\in C(\overline{Q})$ satisfying $u_x\in C(\overline{Q})$. The space X is a Banach space equipped with the norm $\|u\|_X=\max(|u|_{C(\overline{\Omega})},|u_x|_{C(\overline{\Omega})})$.

We shall find a solution in X by a method of continuity which is a version of a fixed point argument [9, Theorem 11.6]. For $\sigma\in[0,1]$ we define a mapping $\mathfrak{F}_\sigma: X\to Y$ by

$$\mathfrak{F}_\sigma(\phi)=\left(\sigma F(\phi_x^2+\gamma^2)^{1/2},\sigma F\max\left\{\alpha,\left(([\nu\phi_x\mathrm{Sign}F]_-)^2+\gamma^2\right)^{1/2}\right\}\Big|_{\partial\Omega}-\sigma\varepsilon\nu\phi_x\Big|_{\partial\Omega}\right),$$

where $Y=C(\overline{Q})\times C(\partial\Omega\times[0,T])$. Let $\mathfrak{H}_\sigma$ denote the solution operator of the problem

$$\begin{cases} u_t-\varepsilon u_{xx}=f & \text{in} \quad Q,\\ u_t=g & \text{on} \quad \partial\Omega\times(0,T),\\ u|_{t=0}=\sigma a & \text{on} \quad \Omega. \end{cases}$$

In other words, it is formally defined by $\mathfrak{H}_\sigma(f,g)=u$ for $(f,g)\in Y$. Let us give a rigorous definition. We replace the boundary condition by the standard Dirichlet condition

$$u(x,t)=\int_0^t g(x,\tau)d\tau+\sigma a(x)=:h(x,t).$$

We extend h linearly in x, i.e.,

$$\tilde{h}(x,t)=\frac{L-x}{L}h(0,t)+\frac{x}{L}h(L,t),\quad x\in\Omega,$$

when $\Omega=(0,L)$. Then $v=u-\tilde{h}$ solves

$$\begin{cases} v_t-\varepsilon v_{xx}=f-\tilde{h}_t & \text{in} \quad Q,\\ v=0 & \text{on} \quad \partial\Omega\times(0,T),\\ v|_{t=0}=\sigma a-\tilde{h}|_{t=0} & \text{on} \quad \Omega. \end{cases}$$

Since $\tilde{h}_t \in C(\overline{Q})$, by the standard L^p theory [11] there is a unique solution v of the above problem, and it belongs to the Sobolev space $W_p^{2,1}(Q)$ for every $p > 1$ if a is sufficiently regular, say $a \in C^2(\overline{\Omega})$. The value $\mathfrak{H}_\sigma(f, g)$ is defined by $v + \tilde{h}$. By the construction the mapping $\mathfrak{H}_\sigma$ is bounded linearly from Y to $W_p^{2,1}(Q)$ for every $p > 1$.

Thus the mapping $\mathfrak{H} : Y \times [0, 1] \to X$ defined by $\mathfrak{H}((f, g), \sigma) = \mathfrak{H}_\sigma(f, g)$ is well-defined and compact by the standard embedding theory [11]. We define $\mathfrak{F} : X \times [0, 1] \to X$ by

$$\mathfrak{F}(\phi, \sigma) = \mathfrak{H}(\mathfrak{F}_\sigma(\phi), \sigma).$$

Evidently, $\mathfrak{F}(\phi, 0) = 0$ for all $\phi \in X$. Moreover $\mathfrak{F}$ is compact since $\mathfrak{H}$ is compact and $\mathfrak{F}$ is continuous. To apply the Leray–Schauder fixed point theorem [9, Theorem 11.6], it remains to prove the a priori estimate

$$\|\phi\|_X < M \quad \text{for } \phi = \mathfrak{F}(\phi, \sigma) \tag{A.1}$$

with M independent of ϕ and σ. We first observe that $\phi \in C^{2,1}(\overline{Q}) \cap C^\infty(Q)$. Since $\phi \in W_p^{2,1}(Q)$ for $p > n + 2$, a standard embedding result [11, Chapter II, Lemma 3.3] implies that $\phi_x \in C^{\mu,\mu/2}(\overline{Q})$ with some $\mu \in (0, 1)$; i.e., ϕ_x is Hölder continuous in $\overline{Q}$. This implies that $\mathfrak{F}_\sigma(\phi) \in C^{\mu,\mu/2}(\overline{Q}) \times C^{\mu/2}(\partial\Omega \times [0, T])$. Since $\mathfrak{F}(\phi, \sigma) = \phi$ and $a \in C^{2+\mu}(\overline{\Omega})$, by the Schauder estimates [11] we conclude that $\phi \in C^{2+\mu,1+\mu/2}(\overline{Q})$. Since $F \in C^\infty(Q)$, then a standard bootstrap argument [11] yields $\phi \in C^\infty(Q)$. Thus the estimate (A.1) is obtained in Corollary 4.3. Note that the term $\varepsilon\nu u_x$ plays a crucial role here. We have thus proved that there exists $u \in X$ such that $\mathfrak{F}(u, 1) = u$, which is the desired solution. So, Theorem A.1 has been proved.

Remark 3. Of course there is another way to prove Theorem A.1. A local-in-time classical solution u can be constructed as in [7]. Once there is a bound for u_x, then the solution can be extended globally in time as in [8]. However, Theorem A.1 is not explicitly included in these references, so we have given a complete proof for the reader's convenience. The solution in Theorem A.1 is actually unique, although we do not use this property.

Acknowledgments. This work was carried out during extremely enjoyable and stimulating visits by C. M. E. to Hokkaido University, whose hospitality he gratefully acknowledges. We are grateful to the referees for their valuable comments.

REFERENCES

[1] S. Angenent and M. E. Gurtin, *General contact-angle conditions with and without kinetics*, Quart. Appl. Math., 54 (1996), pp 557–569.

[2] G. Barles, *Solutions de viscosité des équations de Hamilton-Jacobi*, Springer-Verlag, New York, 1994.

[3] S. J. Chapman, *A mean-field model of superconducting vortices in three dimensions*, SIAM J. Appl. Math., 55 (1995), pp. 1259–1274.

[4] J. R. Claisse, *Vortex Density Motion in a Cylindrical Type* II *Superconductor Subject to a Transverse Applied Magnetic Field*, D. Phil. thesis, University of Sussex, Great Britain, 2001.

[5] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[6] C. M. Elliott, R. Schätzle, and B. E. E. Stoth, *Viscosity solutions of a degenerate parabolic elliptic system arising in the mean field theory of superconductivity*, Arch. Ration. Mech. Anal., 145 (1998), pp. 99–127.

[7] J. Escher, *Quasilinear parabolic systems with dynamic boundary conditions*, Comm. Partial Differential Equations, 18 (1993), pp. 1309–1364.

[8] J. Escher, *On the qualitative behaviour of some semilinear parabolic problems*, Differential Integral Equations, 8 (1995), pp. 247–267.
[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1983.
[10] T. Hinterman, *Evolution equations with dynamic boundary conditions*, Proc. Roy. Soc. Edinburgh Sect. A, 113 (1989), pp. 43–60.
[11] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'tseva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

SIAM J. MATH. ANAL.
Vol. 34, No. 4, pp. 882–900

RECOVERY OF SMALL INHOMOGENEITIES FROM THE SCATTERING AMPLITUDE AT A FIXED FREQUENCY*

HABIB AMMARI†, EKATERINA IAKOVLEVA†, AND SHARI MOSKOW‡

Abstract. We rigorously derive the leading order term in the asymptotic expansion of the scattering amplitude of a collection of a finite number of dielectric inhomogeneities of small diameter. We then apply this asymptotic formula for the purpose of identifying the location and certain properties of the shapes of the small inhomogeneities from scattering amplitude measurements at a fixed frequency. Our main idea is to reduce this reconstruction problem to the calculation of an inverse Fourier transform.

Key words. inverse scattering problem, scattering amplitude, Helmholtz equation, dielectric imperfections, reconstruction

AMS subject classifications. 35R30, 78A46

PII. S0036141001392785

1. Introduction. In this paper, we consider three-dimensional electromagnetic scattering from a collection of small dielectric inhomogeneities. We suppose that there is a finite number of dielectric imperfections in $\mathbf{R}^3$, each of the form $z_j + \alpha B_j$, where $B_j \subset \mathbf{R}^3$ is a bounded, smooth (C^∞) domain containing the origin. This regularity assumption could be considerably weakened. The total collection of imperfections thus takes the form

$$\mathcal{I}_\alpha = \cup_{j=1}^m (z_j + \alpha B_j).$$

The points $z_j \in \mathbf{R}^3$, $j = 1, \ldots, m$, that determine the location of the imperfections are assumed to satisfy

$$0 < d_0 \leq |z_j - z_l| \quad \forall\, j \neq l. \tag{1}$$

We also assume that $\alpha > 0$, the common order of magnitude of the diameters of the imperfections, is small enough such that the imperfections are disjoint.

Our first goal is to provide a rigorous derivation of the asymptotic expansion of the scattering amplitude for such a collection of small dielectric imperfections. Our second goal is to use this expansion for efficiently determining the locations and/or shapes of the small inhomogeneities from scattering amplitude measurements at a fixed frequency by reducing the reconstruction problem of the small inhomogeneities to the calculation of an inverse Fourier transform. We expect that our asymptotic formulas will form the basis for very effective computational identification algorithms, aimed at determining information about the small inhomogeneities from scattering amplitude measurements.

*Received by the editors July 22, 2001; accepted for publication (in revised form) August 9, 2002; published electronically March 5, 2003. This work was partially supported by ACI Jeunes Chercheurs (0693) from the Ministry of Education and Scientific Research, France, and the National Science Foundation DMS1613568-12.

http://www.siam.org/journals/sima/34-4/39278.html

†Centre de Mathématiques Appliquées, CNRS UMR 7641 and Ecole Polytechnique, 91128 Palaiseau Cedex, France (ammari@cmapx.polytechnique.fr, iakov@cmapx.polytechnique.fr).

‡Department of Mathematics, University of Florida, Gainesville, FL 32611 (moskow@math.ufl.edu).

To the best of our knowledge, the present paper is the first attempt to design an effective and accurate method to determine the location and the size of small dielectric inhomogeneities with *both* different electric permittivities and magnetic permeabilities from scattering amplitude measurements. Our method is quite similar to the ideas used by Calderon [5] in his proof of uniqueness of the linearized conductivity problem and later by Sylvester and Uhlmann in their important work [20] on uniqueness of the three-dimensional inverse conductivity problem. Our technique for studying the scattering problem is to reduce the problem to a bounded domain with the aid of integral equation methods. On the bounded domain, the derivation of the asymptotic expansion of the solution relies heavily on the results of [24]. The current work is also a natural extension of the identification procedure that we have presented in [2], where we demonstrated numerically its accuracy and stability. For discussions on other closely related inverse scattering problems, the reader is referred, for example, to [7], [15], [10], [11], [12], [22], [23], [14], [17], [18], [19], and [8].

Let $\mu^0 > 0$ and $\varepsilon^0 > 0$ denote the permeability and the permittivity of the free space; we shall assume that these are positive constants. Let $\mu^j > 0$ and $\varepsilon^j > 0$ denote the permeability and the permittivity of the jth inhomogeneity, $z_j + \alpha B_j$; these are also assumed to be positive constants. Using this notation, we introduce the piecewise constant magnetic permeability

$$\mu_\alpha(x) = \begin{cases} \mu^0, & x \in \mathbf{R}^3 \setminus \bar{\mathcal{I}}_\alpha, \\ \mu^j, & x \in z_j + \alpha B_j,\ j = 1, \dots, m. \end{cases} \tag{2}$$

If we allow the degenerate case $\alpha = 0$, then the function $\mu_0(x)$ equals the constant μ^0. The piecewise constant electric permittivity $\varepsilon_\alpha(x)$ is defined analogously. We need to introduce some additional notation. Let γ^j, $1 \le j \le m$, be a set of positive constants. In effect, $\{\gamma^j\}$ will be either the set $\{\varepsilon^j\}$ or the set $\{\mu^j\}$. For any fixed $1 \le j_0 \le m$, let γ denote the coefficient given by

$$\gamma(x) = \begin{cases} \gamma^0, & x \in \mathbf{R}^3 \setminus \bar{B}_{j_0}, \\ \gamma^{j_0}, & x \in B_{j_0}. \end{cases} \tag{3}$$

By ϕ_l, $1 \le l \le 3$, we denote the solution to

$$\begin{aligned} &\nabla_y \cdot \gamma(y) \nabla_y \phi_l = 0 \quad \text{in } \mathbf{R}^3, \\ &\phi_l - y_l \to 0 \quad \text{as } |y| \to \infty. \end{aligned}$$

This problem may alternatively be written as

$$\begin{cases} \Delta \phi_l = 0 \quad \text{in } B_{j_0}, \text{ and in } \mathbf{R}^3 \setminus \overline{B_{j_0}}, \\ \phi_l \text{ is continuous across } \partial B_{j_0}, \\ \dfrac{\gamma^0}{\gamma^{j_0}} (\partial_\nu \phi_l)^+ - (\partial_\nu \phi_l)^- = 0 \quad \text{on} \quad \partial B_{j_0}, \\ \phi_l(y) - y_l \to 0 \quad \text{as } |y| \to \infty. \end{cases}$$

Here ν denotes the outward unit normal to $\partial(z_j + \alpha B_j)$; superscripts $+$ and $-$ indicate the limiting values as we approach $\partial(z_j + \alpha B_j)$ from outside $z_j + \alpha B_j$ and from inside $z_j + \alpha B_j$. It is obvious that the function ϕ_l depends only on the coefficients γ^0 and γ^{j_0} through the ratio $c = \frac{\gamma^0}{\gamma^{j_0}}$. The existence and uniqueness of this ϕ_l can be established using single layer potentials with suitably chosen densities. It is essential here that

the constant c, by assumption, cannot be 0 or a negative real number. We now define the polarization tensor $M^{j_0}(c)$ of the inhomogeneity B_{j_0} (with aspect ratio c), by

$$M_{kl}^{j_0}(c) = c^{-1} \int_{B_{j_0}} \partial_{y_k} \phi_l \, dy. \tag{4}$$

It is quite easy to see that the tensor $M_{kl}^{j_0}(c)$ is symmetric; since c is a positive real number, it is furthermore positive definite (see [6], [13]).

2. Asymptotic formula for the solution. Consider in this section a homogeneous background medium in all of $\mathbf{R}^3$ with electric permittivity ε^0 and magnetic permeability μ^0, and let ε_α and μ_α be the corresponding dielectric functions in the presence of the small inhomogeneities described above. Let u_α be the solution to the Helmholtz equation

$$\left(\nabla \cdot \frac{1}{\mu_\alpha} \nabla + \omega^2 \varepsilon_\alpha \right) u_\alpha = 0 \quad \text{in} \quad \mathbf{R}^3, \tag{5}$$

with the radiation condition as $r \to \infty$,

$$|\partial_r (u_\alpha - e^{ik\eta \cdot x}) - ik(u_\alpha - e^{ik\eta \cdot x})| = O\left(\frac{1}{r^2} \right), \tag{6}$$

where ω is the frequency, $k^2 = \omega^2 \varepsilon^0 \mu^0$, η is a vector on the unit sphere S^2 in $\mathbf{R}^3$, $\eta \cdot \eta = 1$, and $u_0 = e^{ik\eta \cdot x}$ is an incident plane wave. Note that u_0 satisfies the homogeneous Helmholtz equation

$$\left(\nabla \cdot \frac{1}{\mu^0} \nabla + \omega^2 \varepsilon^0 \right) u_0 = 0 \quad \text{in} \quad \mathbf{R}^3. \tag{7}$$

In this section, we find and prove a formula, asymptotic with respect to the inhomogeneity size α, for u_α in terms of u_0. We begin by defining the outgoing Green function $G(x, y)$ to satisfy

$$\begin{aligned} \left(\Delta_y + k^2 \right) G(x, y) &= -\delta_x(y) \quad \text{in} \quad \mathbf{R}^3, \\ |\partial_r G - ikG| &= O\left(\frac{1}{r^2} \right) \quad \text{as} \quad r \to \infty. \end{aligned} \tag{8}$$

In fact, we know G explicitly:

$$G(x, y) = \frac{e^{ik|x-y|}}{4\pi |x - y|}.$$

Let Ω denote some fixed domain in $\mathbf{R}^3$ that contains the inhomogeneities. Without loss of generality, we can assume that k^2 is not an eigenvalue of $-\Delta$ in Ω corresponding to Dirichlet boundary conditions on $\partial\Omega$. We know that Proposition 1 in [24], which is based on properties of collectively compact operators, guarantees that, for α sufficiently small, the trivial solution is the unique solution to $(\nabla \cdot \frac{1}{\mu_\alpha} \nabla + \omega^2 \varepsilon_\alpha) v_\alpha = 0$ in Ω, with the boundary condition $v_\alpha = 0$ on $\partial\Omega$.

If we consider the equation for u_α in the *exterior* of Ω, multiply G, and integrate by parts, we get that, for $x \in \mathbf{R}^3 \setminus \overline{\Omega}$,

$$u_\alpha(x) = u_0(x) + \int_{\partial\Omega} \partial_{\nu_y} G u_\alpha(y) \, d\sigma_y - \int_{\partial\Omega} G \partial_\nu u_\alpha(y) \, d\sigma_y,$$

where ν is the unit outward normal to $\partial\Omega$. Of course, this equation does not hold up to the boundary of Ω, but if we take the limit as $x \to \partial\Omega$, we get (see, for example, [7] and [16])

$$\frac{1}{2}u_\alpha|_{\partial\Omega} = u_0|_{\partial\Omega} + \int_{\partial\Omega} \partial_{\nu_y} G u_\alpha(y)\, d\sigma_y - \int_{\partial\Omega} G \partial_\nu u_\alpha(y)\, d\sigma_y \tag{9}$$

for $x \in \partial\Omega$. Now define the Dirichlet to Neumann map

$$N_\alpha : H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega),$$

$$N_\alpha(f) = \partial_\nu v_\alpha,$$

where v_α is the solution to

$$\begin{aligned} \left(\nabla \cdot \frac{1}{\mu_\alpha}\nabla + \omega^2 \varepsilon_\alpha\right) v_\alpha = 0 \quad &\text{in} \quad \Omega, \\ v_\alpha = f \quad &\text{on} \quad \partial\Omega. \end{aligned} \tag{10}$$

Hence

$$N_\alpha(u_\alpha|_{\partial\Omega}) = \partial_\nu u_\alpha|_{\partial\Omega}.$$

Similarly, let

$$N_0 : H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega)$$

be the Neumann to Dirichlet map for the limiting problem so that

$$N_0(u_0|_{\partial\Omega}) = \partial_\nu u_0|_{\partial\Omega}.$$

We also define the single and double layer potential operators

$$S : H^{-1/2}(\partial\Omega) \to H^{1/2}(\partial\Omega)$$

and

$$D : H^{1/2}(\partial\Omega) \to H^{1/2}(\partial\Omega),$$

where

$$S : g \to \int_{\partial\Omega} G(x,y) g(y)\, d\sigma_y$$

and

$$D : f \to \int_{\partial\Omega} \partial_{\nu_y} G(x,y) f(y)\, d\sigma_y.$$

Using this operator notation, we see that from (9) we have

$$\left(\frac{I}{2} - D + S N_\alpha\right)(u_\alpha|_{\partial\Omega}) = u_0|_{\partial\Omega}.$$

Similarly, u_0 satisfies

$$\left(\frac{I}{2} - D + SN_0\right)(u_0|_{\partial\Omega}) = u_0|_{\partial\Omega}.$$

Define

$$T_\alpha : H^{1/2}(\partial\Omega) \to H^{1/2}(\partial\Omega)$$

by

$$T_\alpha = \frac{I}{2} - D + SN_\alpha, \tag{11}$$

and let

$$T_0 = \frac{I}{2} - D + SN_0. \tag{12}$$

By subtracting the two above equations, we have that

$$T_\alpha(u_\alpha|_{\partial\Omega}) - T_0(u_0|_{\partial\Omega}) = 0,$$

and hence

$$T_\alpha((u_\alpha - u_0)|_{\partial\Omega}) = S(N_0 - N_\alpha)(u_0|_{\partial\Omega}).$$

We will need the following proposition. The reader is referred to the appendix for its proof. In the following proposition and in the remainder of this paper, all asymptotic terms and constants may depend on the separation d_0 of the inhomogeneities.

PROPOSITION 1. *Let T_α be defined by* (11) *and T_0 by* (12). *Then we have the following:*

(a) *T_α converges to T_0 pointwise.*
(b) *$T_\alpha - T_0$ is collectively compact.*
(c) *There exists a constant C that is independent of α and the set of points $\{z_j\}_{j=1}^m$ such that, for any $f \in H^{1/2}(\partial\Omega)$, T_α^{-1} exists and*

$$\|T_\alpha^{-1} f\|_{H^{1/2}(\partial\Omega)} \leq C\|f\|_{H^{1/2}(\partial\Omega)}.$$

(d) *The following asymptotic formula holds:*

$$\begin{aligned}(T_0 - T_\alpha)(u_0|_{\partial\Omega})(x) &= S(N_0 - N_\alpha)(u_0|_{\partial\Omega})(x) \\ &= \alpha^3 \left[\sum_{j=1}^m \left(1 - \frac{\mu^j}{\mu^0}\right) \nabla u_0(z_j) \cdot M^j \left(\frac{\mu^j}{\mu^0}\right) \nabla_y G(x, z_j) \right. \\ &\qquad \left. + k^2 \left(1 - \frac{\varepsilon^j}{\varepsilon^0}\right) u_0(z_j) G(x, z_j)\right] + o(\alpha^3), \end{aligned} \tag{13}$$

where the asymptotic term $o(\alpha^3)$ is independent of $x \in \partial\Omega$ and the set of points $\{z_j\}_{j=1}^m$.

Define the correction

$$u^{(1)}(x) = \sum_{j=1}^m \left(1 - \frac{\mu^j}{\mu^0}\right) \nabla u_0(z_j) \cdot M^j \left(\frac{\mu^j}{\mu^0}\right) \nabla_y G(x, z_j) + k^2 \left(1 - \frac{\varepsilon^j}{\varepsilon^0}\right) u_0(z_j) G(x, z_j) \tag{14}$$

for $x \neq z_j$, $j = 1, \ldots, m$. We have therefore shown that

$$T_\alpha((u_\alpha - u_0)|_{\partial\Omega}) = \alpha^3 u^{(1)}|_{\partial\Omega} + o(\alpha^3) \tag{15}$$

uniformly for $x \in \partial\Omega$. Note that, from the definition of G, $u^{(1)}$ satisfies

$$(\Delta + k^2)u^{(1)} = \sum_{j=1}^{m} \left(1 - \frac{\mu^j}{\mu^0}\right) \nabla u_0(z_j) \cdot M^j \left(\frac{\mu^j}{\mu^0}\right) \nabla \delta_{z_j} + k^2 \left(1 - \frac{\varepsilon^j}{\varepsilon^0}\right) u_0(z_j) \delta_{z_j}, \tag{16}$$

in the sense of distributions, where δ_{z_j} is the Dirac delta function at the point z_j.

LEMMA 1. *Let the correction term $u^{(1)}$ be defined by* (14). *Then we have*

$$T_0(u^{(1)}|_{\partial\Omega}) = u^{(1)}|_{\partial\Omega}.$$

Proof. Multiplying (16) by G, integrating by parts over Ω, and taking the limit as $x \to \partial\Omega$, we get

$$\frac{1}{2} u^{(1)}|_{\partial\Omega} - \int_{\partial\Omega} \partial_{\nu_y} G u^{(1)}(y)\, d\sigma_y + \int_{\partial\Omega} G \partial_\nu u^{(1)}(y)\, d\sigma_y = 0$$

for $x \in \partial\Omega$. Define $v^{(1)}$ as the unique solution to

$$\begin{cases} \Delta v^{(1)} + k^2 v^{(1)} = 0 & \text{in} \quad \Omega, \\ v^{(1)} = u^{(1)} & \text{on} \quad \partial\Omega; \end{cases}$$

that is,

$$\partial_\nu v^{(1)} = N_0(u^{(1)}|_{\partial\Omega}).$$

Green's formula yields, for any $x \in \Omega$ away from the centers of the inhomogeneities,

$$\begin{aligned} \int_{\partial\Omega} G(x,y) \partial_\nu (u^{(1)} - v^{(1)})(y)\, d\sigma_y &= \sum_{j=1}^{m} \left(1 - \frac{\mu^j}{\mu^0}\right) \nabla u_0(z_j) \cdot M^j \left(\frac{\mu^j}{\mu^0}\right) \nabla_y G(x, z_j) \\ &\quad + k^2 \left(1 - \frac{\varepsilon^j}{\varepsilon^0}\right) u_0(z_j) G(x, z_j) - u^{(1)}(x) + v^{(1)}(x) \\ &= v^{(1)}(x). \end{aligned}$$

Hence, for $x \in \partial\Omega$,

$$\int_{\partial\Omega} G(x,y) \partial_\nu (u^{(1)} - v^{(1)})(y)\, d\sigma_y = u^{(1)}(x).$$

Using this, we can rewrite

$$\begin{aligned} \int_{\partial\Omega} G \partial_\nu u^{(1)}(y)\, d\sigma_y &= \int_{\partial\Omega} G N_0(u^{(1)})(y)\, d\sigma_y + \int_{\partial\Omega} G(\partial_\nu u^{(1)}(y) - N_0(u^{(1)})(y))\, d\sigma_y \\ &= \int_{\partial\Omega} G N_0(u^{(1)})(y)\, d\sigma_y + u^{(1)}(x), \end{aligned}$$

from which it follows that

$$\frac{1}{2} u^{(1)}|_{\partial\Omega} - \int_{\partial\Omega} \partial_{\nu_y} G u^{(1)}(y)\, d\sigma_y + \int_{\partial\Omega} G N_0(u^{(1)})(y)\, d\sigma_y = u^{(1)}(x)$$

for $x \in \partial\Omega$. This just says exactly that $T_0(u^{(1)}|_{\partial\Omega}) = u^{(1)}|_{\partial\Omega}$. ∎

LEMMA 2. *The following estimate holds:*

$$\|u_\alpha - u_0 - \alpha^3 u^{(1)}\|_{H^{1/2}(\partial\Omega)} = o(\alpha^3), \tag{17}$$

where the term $o(\alpha^3)$ goes to zero faster than α^3 independent of the set of points $\{z_j\}_{j=1}^m$.

Proof. From (15) it follows that

$$T_\alpha((u_\alpha - u_0 - \alpha^3 u^{(1)})|_{\partial\Omega}) = \alpha^3 u^{(1)}|_{\partial\Omega} - \alpha^3 T_\alpha(u^{(1)}|_{\partial\Omega}) + o(\alpha^3).$$

Lemma 1 yields

$$T_\alpha((u_\alpha - u_0 - \alpha^3 u^{(1)})|_{\partial\Omega}) = \alpha^3 (T_0 - T_\alpha)(u^{(1)}|_{\partial\Omega}) + o(\alpha^3).$$

Therefore, due to the pointwise convergence of T_α to T_0, we obtain

$$T_\alpha((u_\alpha - u_0 - \alpha^3 u^{(1)})|_{\partial\Omega}) = o(\alpha^3),$$

which leads, by using point (c) in Proposition 1, to the desired estimate (17). ∎

From this lemma, we obtain the following theorem.

THEOREM 1. *Let u_α be the solution to* (5)*, and let $M^j(\frac{\mu^j}{\mu^0})$ be the polarization tensors for the shapes B_j defined by* (4)*. Then, for $x \in \mathbf{R}^3 \setminus \overline{\Omega}$ bounded away from $\partial\Omega$, we have the pointwise expansion*

$$\begin{aligned} u_\alpha(x) = {} & e^{ik\eta\cdot x} \\ & + \alpha^3 \sum_{j=1}^m e^{ik\eta\cdot z_j} \left[ik \left(1 - \frac{\mu^j}{\mu^0}\right) \nabla_y G(x, z_j) \cdot M^j \left(\frac{\mu^j}{\mu^0}\right) \eta \right. \\ & \qquad \left. + k^2 \left(1 - \frac{\varepsilon^j}{\varepsilon^0}\right) |B_j| G(x, z_j) \right] \\ & + o(\alpha^3). \end{aligned} \tag{18}$$

Here the remainder term $o(\alpha^3)$ is independent of x and the set of points $\{z_j\}_{j=1}^m$.

Proof. From Lemma 2, it follows that $u_\alpha - u_0$ satisfies in $\mathbf{R}^3 \setminus \overline{\Omega}$

$$\begin{cases} \Delta(u_\alpha - u_0) + k^2(u_\alpha - u_0) = 0 \quad \text{in} \quad \mathbf{R}^3 \setminus \overline{\Omega}, \\ (u_\alpha - u_0) = \alpha^3 u^{(1)} + o(\alpha^3) \quad \text{on} \quad \partial\Omega, \\ |\partial_r(u_\alpha - u_0) - ik(u_\alpha - u_0)| = O(\frac{1}{r^2}). \end{cases}$$

Let $\mathcal{G}$ denote the outgoing Dirichlet Green function that is defined by

$$\begin{cases} \Delta\mathcal{G} + k^2\mathcal{G} = -\delta \quad \text{in} \quad \mathbf{R}^3 \setminus \overline{\Omega}, \\ \mathcal{G} = 0 \quad \text{on} \quad \partial\Omega, \\ |\partial_r\mathcal{G} - ik\mathcal{G}| = O(\frac{1}{r^2}). \end{cases}$$

It is easy to see that $u_\alpha - u_0$ has the following integral representation in $\mathbf{R}^3 \setminus \overline{\Omega}$:

$$(u_\alpha - u_0)(x) = \int_{\partial\Omega} \frac{\partial\mathcal{G}}{\partial\nu_y}(x, y)(u_\alpha - u_0)(y)\, d\sigma(y) \quad \forall\, x \in \mathbf{R}^3 \setminus \overline{\Omega}.$$

Moreover, for any $x \in \mathbf{R}^3 \setminus \overline{\Omega}$ which is bounded away from $\partial\Omega$, we obtain from the asymptotic expansion of the boundary condition in Lemma 2 that

$$(u_\alpha - u_0)(x) = \alpha^3 \int_{\partial\Omega} \frac{\partial \mathcal{G}}{\partial \nu_y}(x,y) u^{(1)}(y)\, d\sigma(y) + o(\alpha^3),$$

where $o(\alpha^3)$ is independent of x and the set of points $\{z_j\}_{j=1}^m$. Since, for any $x \in \mathbf{R}^3 \setminus \overline{\Omega}$ and $z \in \Omega$, we have by standard integration by parts the identities

$$\int_{\partial\Omega} \frac{\partial \mathcal{G}}{\partial \nu}(x,y) G(y,z)\, d\sigma(y) = G(x,z)$$

and

$$\int_{\partial\Omega} \frac{\partial \mathcal{G}}{\partial \nu}(x,y) \nabla_z G(y,z)\, d\sigma(y) = \nabla_z G(x,z),$$

the expression of the correction term $u^{(1)}$ immediately leads to the promised asymptotic expansion. □

We also can obtain the next proposition on the norm convergence of the solutions.

PROPOSITION 2. *There exists a constant C that is independent of α and the set of points $\{z_j\}_{j=1}^m$ such that the following energy estimate holds:*

$$\|u_\alpha - u_0\|_{L^2(\Omega)} + \|\nabla u_\alpha - \nabla u_0\|_{L^2(\Omega)} \le C\alpha^2. \tag{19}$$

Proof. Let $\tilde{u}_\alpha$ be defined as the unique solution to

$$\begin{cases} \Delta \tilde{u}_\alpha + k^2 \tilde{u}_\alpha = 0 & \text{in} \quad \Omega, \\ \tilde{u}_\alpha = u_\alpha & \text{on} \quad \partial\Omega. \end{cases}$$

We have

$$\begin{cases} \Delta(\tilde{u}_\alpha - u_0) + k^2(\tilde{u}_\alpha - u_0) = 0 & \text{in} \quad \Omega, \\ (\tilde{u}_\alpha - u_0) = u_\alpha - u_0 & \text{on} \quad \partial\Omega, \end{cases}$$

which leads to

$$\|\tilde{u}_\alpha - u_0\|_{H^1(\Omega)} \le C \|u_\alpha - u_0\|_{H^{1/2}(\Omega)},$$

where the constant C is independent of α. Using Lemma 2, we get that $\|\tilde{u}_\alpha - u_0\|_{H^1(\Omega)}$ is of order α^3. Now note that the function $(u_\alpha - \tilde{u}_\alpha)$ is in $H_0^1(\Omega)$, and for any $v \in H_0^1(\Omega)$

$$\begin{aligned} \int_\Omega \frac{1}{\mu_\alpha} \nabla(u_\alpha - \tilde{u}_\alpha) \cdot \nabla v - \omega^2 \int_\Omega \varepsilon_\alpha (u_\alpha - \tilde{u}_\alpha) v \quad &= \quad \int_\Omega \frac{1}{\mu_\alpha} \nabla u_\alpha \cdot \nabla v - \omega^2 \int_\Omega \varepsilon_\alpha u_\alpha v \\ &- \quad \int_\Omega \frac{1}{\mu^0} \nabla \tilde{u}_\alpha \cdot \nabla v + \omega^2 \int_\Omega \varepsilon^0 \tilde{u}_\alpha v \\ &+ \quad \sum_{j=1}^m \left(\frac{1}{\mu^0} - \frac{1}{\mu^j} \right) \int_{z_j + \alpha B_j} \nabla \tilde{u}_\alpha \cdot \nabla v \\ &+ \quad k^2 \left(\frac{\varepsilon^j}{\varepsilon^0} - 1 \right) \int_{z_j + \alpha B_j} \tilde{u}_\alpha v. \end{aligned}$$

Next we can bound

$$\left| \int_{z_j + \alpha B_j} \nabla \tilde{u}_\alpha \cdot \nabla v \right| \le \|\nabla \tilde{u}_\alpha\|_{L^2(z_j + \alpha B_j)} \|\nabla v\|_{L^2(\Omega)}$$

and

$$\left| \int_{z_j+\alpha B_j} \tilde{u}_\alpha v \right| \leq \|\tilde{u}_\alpha\|_{L^2(z_j+\alpha B_j)} \|v\|_{L^2(\Omega)}.$$

However, using the triangle inequality,

$$\|\nabla \tilde{u}_\alpha\|_{L^2(z_j+\alpha B_j)} \leq \|\nabla(\tilde{u}_\alpha - u_0)\|_{L^2(\Omega)} + \|\nabla u_0\|_{L^2(z_j+\alpha B_j)},$$

and

$$\|\tilde{u}_\alpha\|_{L^2(z_j+\alpha B_j)} \leq \|(\tilde{u}_\alpha - u_0)\|_{L^2(\Omega)} + \|u_0\|_{L^2(z_j+\alpha B_j)}.$$

Therefore, since

$$\|u_0\|_{H^1(z_j+\alpha B_j)} = O(\alpha^2)$$

and

$$\|(\tilde{u}_\alpha - u_0)\|_{H^1(\Omega)} = O(\alpha^3),$$

we obtain

$$\left| \int_\Omega \frac{1}{\mu_\alpha} \nabla(u_\alpha - \tilde{u}_\alpha) \cdot \nabla v - \omega^2 \int_\Omega \varepsilon_\alpha (u_\alpha - \tilde{u}_\alpha) v \right| \leq C\alpha^2 \|v\|_{H^1(\Omega)}$$

for any $v \in H_0^1(\Omega)$. From Proposition 1 in [24], it then follows that

$$\|(u_\alpha - \tilde{u}_\alpha)\|_{H^1(\Omega)} = O(\alpha^2);$$

hence

$$\|(u_\alpha - u_0)\|_{H^1(\Omega)} \leq \|(u_\alpha - \tilde{u}_\alpha)\|_{H^1(\Omega)} + \|(u_0 - \tilde{u}_\alpha)\|_{H^1(\Omega)} \leq C\alpha^2,$$

exactly as desired. ∎

3. Asymptotic formula for the scattering amplitude. We now use the results derived in the previous section to prove an asymptotic formula for the scattering amplitude. The scattering amplitude, $A_\alpha(\frac{x}{|x|}, \eta, k)$, is defined to be a function which satisfies

$$u_\alpha(x) = e^{ik\eta\cdot x} + A_\alpha\left(\frac{x}{|x|}, \eta, k\right) \frac{e^{ik|x|}}{|x|} + o\left(\frac{1}{|x|}\right) \tag{20}$$

as $|x| \to \infty$. Recall that

$$G(x, z_j) = \frac{e^{ik|x-z_j|}}{4\pi|x-z_j|}.$$

One can show from a simple calculation that, as $|x| \to \infty$,

$$G(x, z_j) = \frac{e^{ik|x|}}{|x|} \frac{e^{-ik\frac{x}{|x|}\cdot z_j}}{4\pi} + o\left(\frac{1}{|x|}\right) \tag{21}$$

and

$$\nabla_y G(x,z_j) = \frac{e^{ik|x|}}{|x|}\frac{ikx}{4\pi|x|}e^{-ik\frac{x}{|x|}\cdot z_j} + o\left(\frac{1}{|x|}\right). \tag{22}$$

The following asymptotic formula for the scattering amplitude holds.

THEOREM 2. *The scattering amplitude*

$$A_\alpha\left(\frac{x}{|x|},\eta,k\right) = \frac{\alpha^3k^2}{4\pi}\sum_{j=1}^{m} e^{ik(\eta-\frac{x}{|x|})\cdot z_j}\left[\left(\frac{\mu^j}{\mu^0}-1\right)\frac{x}{|x|}\cdot M^j\eta - \left(\frac{\varepsilon^j}{\varepsilon^0}-1\right)|B_j|\right] + o(\alpha^3) \tag{23}$$

for any $\frac{x}{|x|}$ *and* $\eta \in S^2$, *where* $o(\alpha^3)$ *is independent of the set of points* $\{z_j\}_{j=1}^m$.

Proof. This follows from (21), (22), and the expansion in Theorem 1.

4. Method for reconstruction of inhomogeneities at a fixed frequency. In this section, we present a linear method to determine the locations and the polarization tensors of the small inhomogeneities from scattering amplitude measurements for a fixed frequency. Based on the asymptotic expansion (23), we reduce the reconstruction of the small dielectric inhomogeneities from the scattering amplitude to the calculation of an inverse Fourier transform. For convenience, we are going to assume that B_j, for $j = 1,\dots,m$, are balls. In this case, the polarization tensors M^j have the following explicit forms (see, for example, [25]):

$$M^j\left(\frac{\mu^j}{\mu^0}\right) = m^j I_3,$$

where I_3 is the 3×3 identity matrix and the scalars m^j are given by

$$m^j = 8\pi|B_j|\frac{\mu^j}{\mu^j+\mu^0}.$$

We assume that we are in possession of the scattering amplitude $A_\alpha(\frac{x_l}{|x_l|},\eta_{l'},k)$ for a collection of pairs $(\frac{x_l}{|x_l|},\eta_{l'})$, where $l = 1,\dots,L$ and $l' = 1,\dots,L'$. Introduce

$$g\left(\frac{x}{|x|},\eta\right) = \sum_{j=1}^{m} e^{ik(\eta-\frac{x}{|x|})\cdot z_j}\left[\left(\frac{\mu^j}{\mu^0}-1\right)m^j\frac{x}{|x|}\cdot\eta - \left(\frac{\varepsilon^j}{\varepsilon^0}-1\right)|B_j|\right], \quad \frac{x}{|x|},\ \eta\in S^2. \tag{24}$$

We first observe that

$$g\left(\frac{x}{|x|},\eta\right) = g\left(-\eta,-\frac{x}{|x|}\right) \quad \forall\ \frac{x}{|x|},\eta\in S^2. \tag{25}$$

Define, for $l = 1,\dots,L$ and $l' = 1,\dots,L'$, the coefficients $a_{l,l'}$ by

$$a_{l,l'} = \frac{4\pi}{k^2\alpha^3}A_\alpha\left(\frac{x_l}{|x_l|},\eta_{l'},k\right).$$

Our reconstruction procedure is divided into three steps.

Step 1. Given that

$$g\left(\frac{x_l}{|x_l|},\eta_{l'}\right) \approx a_{l,l'},$$

we can compute using the fast Fourier transform (FFT) an accurate approximation of $g(\frac{x}{|x|},\eta)$ on $S^2 \times S^2$.

Step 2. Let M denote the following complex variety:

$$M = \{\xi \in C^3,\ \xi \cdot \xi = 1\}.$$

It is easy to see that $g(\frac{x}{|x|},\eta)$ has an analytic continuation to $M \times M$. Let $(Y_{p,q})_{-p\le q\le p,\, p=0,1,\ldots,}$ denote the normalized (in $L^2(S^2)$) spherical harmonics. Denote by $g_{p,q}$ the Fourier coefficients of g:

$$g\left(\frac{x}{|x|},\eta\right) = \sum_{p,q} g_{p,q}\left(\frac{x}{|x|}\right) Y_{p,q}(\eta) \quad \forall\ \frac{x}{|x|},\eta \in S^2. \tag{26}$$

Recall that, from Step 1, we are in fact in possession of an accurate approximation of $g_{p,q}(\frac{x}{|x|})$ on S^2 for $-p \le q \le p$ and $p \le P$. In view of (26), the analytic continuation of the truncated Fourier series

$$\sum_{p,q;\, p\le P} g_{p,q}\left(\frac{x}{|x|}\right) Y_{p,q}(\eta)$$

of $g(\frac{x}{|x|},\eta)$ on $M \times M$ can be obtained by using the standard analytic continuation of the spherical harmonics $(Y_{p,q}(\eta))_{p,q}$ on the complex variety M followed by another analytic continuation of the Fourier expansion in $\frac{x}{|x|}$. We know that the analytic continuation of g from $S^2 \times S^2$ to $M \times M$ is unique.

Step 3. Recall that, given $a_{l,l'}$ for $l = 1,\ldots,L$ and $l' = 1,\ldots,L'$, we have constructed by Steps 1 and 2 an accurate approximation of the function $g(\frac{x}{|x|},\eta)$ that is analytic on $M \times M$ and is such that

$$g\left(\frac{x_l}{|x_l|},\eta_{l'}\right) \approx a_{l,l'} \quad \forall\, l = 1,\ldots,L \text{ and } l' = 1,\ldots,L'.$$

However, for any $\xi \in \mathbf{R}^3$, we know that there exist ξ_1 and ξ_2 in M such that $\xi = k(\xi_1 - \xi_2)$; see, for example, [5] and [20]. Let us now view $(a_{l,l'})$ as a function of $\xi \in \mathbf{R}^3$. We have

$$g(\xi_1,\xi_2) = \sum_{j=1}^{m} e^{-i\xi\cdot z_j}\left[\left(\frac{\mu^j}{\mu^0} - 1\right) m^j \xi_1 \cdot \xi_2 - \left(\frac{\varepsilon^j}{\varepsilon^0} - 1\right)|B_j|\right],$$

and, since

$$\xi_1 \cdot \xi_2 = 1 - \frac{1}{2}k^2|\xi|^2,$$

we can rewrite g as follows:

$$g(\xi_1,\xi_2) = \sum_{j=1}^{m} e^{-i\xi\cdot z_j}\left[\left(\frac{\mu^j}{\mu^0} - 1\right) m^j \left(1 - \frac{1}{2}k^2|\xi|^2\right) - \left(\frac{\varepsilon^j}{\varepsilon^0} - 1\right)|B_j|\right]. \tag{27}$$

Define

$$\tilde{g}(\xi) = g(\xi_1,\xi_2),$$

and note that we are now in possession of an approximation to $\tilde{g}(\xi)$ for any $\xi \in \mathbf{R}^3$. Here we rely on the fact that the analytic continuation is unique.

Recall that $e^{-i\xi\cdot z_j}$ (up to a multiplicative constant) is exactly the Fourier transform of the Dirac function δ_{z_j} (a point mass located at z_j). Multiplication by powers of ξ in Fourier space corresponds to differentiation of the Dirac function. Therefore, using the inverse Fourier transform, we obtain

$$\mathcal{F}^{-1}(\tilde{g}(\xi)) = \sum_{j=1}^{m} L_j(\delta_{z_j}),$$

where L_j are, in view of (27), second order constant coefficient differential operators.

Hence $\tilde{g}(\xi)$ is the inverse Fourier transform of a distribution with its support at the locations of the centers of inhomogeneities z_j. Therefore, we think that a numerical Fourier inversion of a sample of $(\tilde{g}(\xi))$ will efficiently pin down the z_j's. The method of location of the points z_j is then similar to that proposed for the conductivity problem [2] from boundary measurements. The number of data (sampling) points needed for an accurate discrete Fourier inversion of $\tilde{g}(\xi)$ follows from the Shannon theorem [9]. We need (conservatively), of order $(h/\delta)^3$, sampled values of ξ to reconstruct, with resolution δ, a collection of inhomogeneities that lie inside a square of side h. Note, however, that real measurements are taken only in Step 1. It remains to be seen how many such measurements are needed. Once the locations $\{z_j\}_{j=1}^m$ are known, we may calculate $|B_j|$ by solving the appropriate linear system arising from (27). If B_j are general domains, our calculations become more complex, and eventually we have to deal with pseudodifferential operators (independent of the space variable x) applied to the same Dirac functions. Numerical experiments examining the feasibility of this approach will be presented in a forthcoming publication.

Appendix. Proof of Proposition 1. Recall that Ω is some fixed domain in $\mathbf{R}^3$ containing the inhomogeneities. Define $\hat{G}(x,z)$ to be the Dirichlet Green function for Ω,

$$\begin{aligned} &\Delta_z \hat{G}(x,z) + k^2 \hat{G}(x,z) = -\delta_x \quad \text{in} \quad \Omega, \\ &\hat{G}(x,z) = 0 \quad \text{on} \quad \partial\Omega. \end{aligned} \tag{28}$$

Recall that

$$N_\alpha f - N_0 f = \frac{\partial v_\alpha}{\partial \nu} - \frac{\partial v_0}{\partial \nu},$$

where

$$\nabla \cdot \frac{1}{\mu_\alpha} \nabla v_\alpha + \omega^2 \varepsilon_\alpha v_\alpha = 0 \quad \text{in} \quad \Omega, \tag{29}$$

$$v_\alpha = f \quad \text{on} \quad \partial\Omega,$$

and

$$\nabla \cdot \frac{1}{\mu^0} \nabla v_0 + \omega^2 \varepsilon^0 v_0 = 0 \quad \text{in} \quad \Omega, \tag{30}$$

$$v_0 = f \quad \text{on} \quad \partial\Omega.$$

Integration by parts gives

$$\begin{aligned} v_\alpha(x) &= -\int_\Omega v_\alpha(z)(\Delta_z \hat{G} + k^2 \hat{G})\, dz \\ &= \int_{\partial\Omega} f \frac{\partial \hat{G}}{\partial \nu_z}\, d\sigma_z + \int_\Omega \nabla v_\alpha \cdot \nabla_z \hat{G}\, dz - \int_\Omega k^2 v_\alpha \hat{G}\, dz \\ &= v_0(x) + \sum_{j=1}^m \int_{z_j+\alpha B_j} \left[\left(1 - \frac{\mu^0}{\mu^j}\right) \nabla v_\alpha \nabla \hat{G} dz + k^2 \left(1 - \frac{\varepsilon^j}{\varepsilon^0}\right) v_\alpha \hat{G} \right], \end{aligned} \tag{31}$$

since by (29) and (30)

$$\int_\Omega \frac{1}{\mu_\alpha} \nabla v_\alpha \cdot \nabla_z \hat{G}\, dz - \omega^2 \int_\Omega \varepsilon_\alpha v_\alpha \hat{G}\, dz = 0$$

and

$$v_0(x) = \int_{\partial\Omega} f \frac{\partial \hat{G}}{\partial \nu_z}\, d\sigma_z.$$

We first derive a uniform asymptotic expansion for $\frac{\partial v_\alpha}{\partial \nu}$ on $\partial\Omega$. We note that this is similar to Theorem 1 in [24], where the authors derived an expansion when $n = 2$ using the free space Green function. We use the Dirichlet Green function because it is more convenient for our purposes.

LEMMA 3. *Let v_α and v_0 be defined as above. Then we have the pointwise expansion*

$$\begin{aligned} &(N_\alpha - N_0)(f) \\ &\quad = \frac{\partial v_\alpha}{\partial \nu}(x) - \frac{\partial v_0}{\partial \nu}(x) \\ &\quad = \alpha^3 \sum_{j=1}^m \left[\left(\frac{1}{\mu^j} - \frac{1}{\mu^0} \right) \nabla v_0(z_j) \cdot M^j \left(\frac{\mu^j}{\mu^0} \right) \nabla_y \frac{\partial}{\partial \nu_x} \hat{G}(x, z_j) \right. \\ &\qquad \left. + k^2 \left(1 - \frac{\varepsilon^j}{\varepsilon^0}\right) v_0(z_j) \frac{\partial}{\partial \nu_x} \hat{G}(x, z_j) \right] + o(\alpha^3), \end{aligned} \tag{32}$$

where the term $o(\alpha^3)$ is uniform for $x \in \partial\Omega$.

For reasons of brevity, we restrict a significant part of the derivation of the asymptotic expansion (32) to the case of one inhomogeneity ($m = 1$). We suppose that this inhomogeneity is centered at the origin, so it is of the form αB. The general case may be verified by a fairly direct iteration of the argument we will present here, adding one inhomogeneity at a time. We will as usual make the change of variables

$$y = x/\alpha,$$

where

$$\tilde{\Omega} = \frac{1}{\alpha}\Omega$$

and

$$B = \frac{1}{\alpha} B_\alpha.$$

Define the correction $w_\alpha(y)$ to be the unique solution to

$$
\begin{aligned}
&\Delta_y w_\alpha + \alpha^2\omega^2\varepsilon^1\mu^1 w_\alpha = 0 \quad \text{in} \quad B,\\
&\Delta_y w_\alpha + \alpha^2\omega^2\varepsilon^0\mu^0 w_\alpha = 0 \quad \text{in} \quad \tilde{\Omega}\setminus\overline{B},\\
&\frac{1}{\mu^0}\frac{\partial w_\alpha}{\partial \nu_y}^+ - \frac{1}{\mu^1}\frac{\partial w_\alpha}{\partial \nu_y}^- = -\left(\frac{1}{\mu^0}-\frac{1}{\mu^1}\right)\nabla_x v_0(0)\cdot\nu \quad \text{on} \quad \partial B,\\
&w_\alpha = 0 \quad \text{on} \quad \partial\tilde{\Omega},
\end{aligned}
\tag{33}
$$

with

$$w_\alpha \text{ continuous across } \partial B.$$

Also, define $w(y)$, which is independent of α and a sort of limit of w_α, as the unique solution to

$$
\begin{aligned}
&\Delta_y w = 0 \quad \text{in} \quad B,\\
&\Delta_y w = 0 \quad \text{in} \quad \mathbf{R}^n\setminus\overline{B},\\
&\frac{1}{\mu^0}\frac{\partial w}{\partial \nu_y}^+ - \frac{1}{\mu^1}\frac{\partial w}{\partial \nu_y}^- = -\left(\frac{1}{\mu^0}-\frac{1}{\mu^1}\right)\nabla_x v_0(0)\cdot\nu \quad \text{on} \quad \partial B,\\
&\lim_{|y|\to\infty}|w(y)| = 0,
\end{aligned}
\tag{34}
$$

with

$$w \text{ continuous across } \partial B.$$

Recall that $|w(y)| = O(\frac{1}{|y|})$ as $|y|\to+\infty$. We now need to prove two lemmas before we can proceed with the derivation of the asymptotic formula (13).

LEMMA 4. *Let v_α, v_0, and w_α be given by* (29), (30), *and* (33), *respectively. Let*

$$z_\alpha(y) = v_\alpha(\alpha y) - v(\alpha y) - \alpha w_\alpha(y).$$

Then there exists a constant C independent of α such that

$$\|z_\alpha\|_{L^2(\tilde{\Omega})} \le C$$

and

$$\|\nabla_y z_\alpha\|_{L^2(\tilde{\Omega})} \le C\alpha.$$

Proof. Note that $z_\alpha(x/\varepsilon)\in H_0^1(\Omega)$. For any $\phi\in H_0^1(\Omega)$, integration by parts gives us that

$$
\begin{aligned}
&\int_{\tilde{\Omega}}\frac{1}{\mu_\alpha}\nabla_y z_\alpha\cdot\nabla_y\phi(\alpha y)\,dy - \alpha^2\omega^2\int_{\tilde{\Omega}}\varepsilon_\alpha(\alpha y) z_\alpha\phi(\alpha y)\,dy\\
&= \left(\frac{1}{\mu^0}-\frac{1}{\mu^1}\right)\int_{\partial B}\nabla_x(v_0(\alpha y)-v_0(0))\cdot\nu\phi(\alpha y)\,d\sigma_y - \alpha^2\omega^2(\varepsilon^0-\varepsilon^1)\int_B v_0(\alpha y)\phi(\alpha y)\,dy\\
&= \left(\frac{1}{\mu^0}-\frac{1}{\mu^1}\right)\int_B \alpha\Delta_x(v_0(\alpha y)-v_0(0))\phi(\alpha y)\,d\sigma_y - \alpha^2\omega^2(\varepsilon^0-\varepsilon^1)\int_B v_0(\alpha y)\phi(\alpha y)\,dy\\
&\quad + \left(\frac{1}{\mu^0}-\frac{1}{\mu^1}\right)\int_B \nabla_x(v_0(\alpha y)-v_0(0)\cdot\nabla_y\phi(\alpha y)\,d\sigma_y.
\end{aligned}
$$

Next we change variables back to the small domain on the left-hand side and multiply by α to obtain

$$\begin{aligned}&\int_\Omega \frac{1}{\mu_\alpha}\nabla_x z_\alpha \cdot \nabla_x \phi\, dx - \omega^2 \int_\Omega \varepsilon_\alpha z_\alpha \phi\, dx \\ &= \alpha^2 \left(\frac{1}{\mu^0} - \frac{1}{\mu^1}\right) \int_B \Delta_x (v_0(\alpha y) - v_0(0))\phi(\alpha y)\, dy - \alpha^3 \omega^2 (\varepsilon^0 - \varepsilon^1) \int_B v_0(\alpha y)\phi(\alpha y)\, dy \\ &\quad + \alpha \left(\frac{1}{\mu^0} - \frac{1}{\mu^1}\right) \int_B \nabla_x (v_0(\alpha y) - v_0(0)) \cdot \nabla_y \phi(\alpha y)\, dy.\end{aligned}$$

Using a Taylor expansion of v_0, we find that there exists C, depending on v_0 but independent of α, such that

$$\left| \int_\Omega \frac{1}{\mu_\alpha}\nabla_x z_\alpha \cdot \nabla_x \phi\, dx - \omega^2 \int_\Omega \varepsilon_\alpha z_\alpha \phi\, dx \right| \leq C\alpha^3 \|\phi(\alpha y)\|_{L^2(B)} + C\alpha^2 \|\nabla_y \phi(\alpha y)\|_{L^2(B)}.$$

By rescaling, we see that

$$\|\phi(\alpha y)\|_{L^2(B)} = \alpha^{-3/2} \|\phi\|_{L^2(\alpha B)}$$

and

$$\|\nabla_y \phi(\alpha y)\|_{L^2(B)} = \alpha^{-1/2} \|\nabla_x \phi\|_{L^2(\alpha B)}$$

so that

$$\left| \int_\Omega \frac{1}{\mu_\alpha}\nabla_x z_\alpha \cdot \nabla_x \phi\, dx - \omega^2 \int_\Omega \varepsilon_\alpha z_\alpha \phi\, dx \right| \leq C\alpha^{3/2} \|\phi\|_{H^1(\Omega)}.$$

By Proposition 1 of [24], it follows that

$$\|z_\alpha\|_{H^1(\Omega)} \leq C\alpha^{3/2}.$$

The result then follows from another scaling. ∎

LEMMA 5. *Let w_α and w be defined by* (33) *and* (34)*, respectively. Then there exists C independent of α such that*

$$\|\nabla_y (w_\alpha - w)\|_{L^2(\tilde{\Omega})} \leq \frac{C}{\alpha^{1/2}}.$$

Proof. Consider $w_\alpha(x/\alpha) - w(x/\alpha)$. Since w_α and w share the same jump condition on the boundary of the ball, their difference satisfies an equation across this boundary. It is not hard to see that in fact we have

$$\begin{aligned}\nabla_x \cdot \frac{1}{\mu_\alpha}\nabla_x (w_\alpha - w) + \omega^2 \varepsilon_\alpha (w_\alpha - w) &= -\omega^2 \varepsilon_\alpha w \quad \text{in} \quad \Omega, \\ w_\alpha - w &= -w \quad \text{on} \quad \partial\Omega.\end{aligned}$$

By Proposition 1 and Corollary 1 in [24], there exists a constant C independent of α such that

$$\|w_\alpha - w\|_{H^1(\Omega)} \leq C \left(\|w\|_{L^2(\Omega)} + \|w\|_{H^{1/2}(\partial\Omega)} \right).$$

Since Ω is a bounded domain and $w(y)$ is bounded, we clearly have $\|w\|_{L^2(\Omega)}$ bounded. Also, since $w(x/\alpha)$ decays as $\alpha \to 0$, we also have $\|w\|_{H^{1/2}(\partial\Omega)}$ bounded independently of α. Hence

$$\|w_\alpha - w\|_{H^1(\Omega)} \leq C,$$

which by rescaling proves the lemma. □

Now define

$$r_\alpha(y) = v_\alpha(\alpha y) - v_0(\alpha y) - \alpha w - c_\alpha,$$

where the constant c_α is defined so that r_α satisfies

$$\int_{\partial B} r_\alpha \, d\sigma_y = 0.$$

The previous two lemmas together imply that

$$\|\nabla_y r_\alpha\|_{L^2(\tilde{\Omega})} \leq C\alpha^{1/2}.$$

Then, from (31),

$$\begin{aligned}
v_\alpha(x) - v_0(x) &= \int_{\alpha B} \left[\left(1 - \frac{\mu^0}{\mu^1}\right) \nabla_z v_\alpha(z) \nabla_z \hat{G}(x,z) + k^2 \left(1 - \frac{\varepsilon^1}{\varepsilon^0}\right) v_\alpha(z) \hat{G}(x,z) \right] dz \\
&= \alpha^3 \int_B \left[\left(1 - \frac{\mu^0}{\mu^1}\right) \nabla_z v_\alpha(\alpha y) \nabla_z \hat{G}(x, \alpha y) \right. \\
&\quad \left. + k^2 \left(1 - \frac{\varepsilon^1}{\varepsilon^0}\right) v_\alpha(\alpha y) \hat{G}(x, \alpha y) \right] dy \\
&= \alpha^2 \int_B \left(1 - \frac{\mu^0}{\mu^1}\right) \nabla_y (v_0 + \alpha w) \nabla_z \hat{G}(x, \alpha y) \\
&\quad + k^2 \alpha^3 \int_B \left(1 - \frac{\varepsilon^1}{\varepsilon^0}\right) v_\alpha(\alpha y) \hat{G}(x, \alpha y) \, dy \\
&\quad + \alpha^2 \int_B \left(1 - \frac{\mu^0}{\mu^1}\right) \nabla_y (r_\alpha) \nabla_z \hat{G}(x, \alpha y) \, dy. \qquad (35)
\end{aligned}$$

By expanding $\hat{G}$ in a Taylor series and using the above estimate for r_α, we have that

$$\int_B \nabla_y r_\alpha \cdot \nabla_x \hat{G}(x, \alpha y) \, dy = \int_B \nabla_y r_\alpha \cdot \nabla_x \hat{G}(x, 0) \, dy + O(\alpha^{3/2}), \qquad (36)$$

and since we have chosen r_α to have integral zero around the boundary of B, the first term on the right-hand side above is zero by integration by parts. Hence

$$\int_B \nabla_y r_\alpha \cdot \nabla_x \hat{G}(x, \alpha y) \, dy = O(\alpha^{3/2}). \qquad (37)$$

Inserting this into (35), we have shown that

$$\begin{aligned}
v_\alpha(x) - v_0(x) &= \alpha^2 \int_B \left(1 - \frac{\mu^0}{\mu^1}\right) \nabla_y (v_0 + \alpha w) \nabla_z \hat{G}(x, \alpha y) \\
&\quad + \alpha^3 k^2 \int_B \left(1 - \frac{\varepsilon^1}{\varepsilon^0}\right) v_\alpha(\alpha y) \hat{G}(x, \alpha y) \, dy + o(\alpha^3). \qquad (38)
\end{aligned}$$

From this expression, we now derive the formulae with the polarization tensor:

$$\begin{aligned} v_\alpha(x) - v_0(x) &= \alpha^3 \left(1 - \frac{\mu^0}{\mu^1}\right) \left[\int_B \nabla_x v_0(\alpha y) \cdot \nabla_z \hat{G}(x, \alpha y)\, dy \right. \\ &\quad \left. + \int_B \nabla_y w \cdot \nabla_z \hat{G}(x, \alpha y)\, dy\right] \\ &\quad + k^2 \alpha^3 \int_B \left(1 - \frac{\varepsilon^1}{\varepsilon^0}\right) v_\alpha(\alpha y) \hat{G}(x, \alpha y)\, dy + o(\alpha^3) \qquad (39) \\ &= \alpha^3 \left(1 - \frac{\mu^0}{\mu^1}\right) |B| \nabla_x v_0(0) \cdot \nabla_z \hat{G}(x, 0) \\ &\quad + \alpha^3 \left(1 - \frac{\mu^0}{\mu^1}\right) \int_B \nabla_y w \cdot \nabla_z \hat{G}(x, 0)\, dy \qquad (40) \\ &\quad + k^2 \alpha^3 \left(1 - \frac{\varepsilon^1}{\varepsilon^0}\right) |B| v_0(0)\, \hat{G}(x, 0) + o(\alpha^3) \qquad (41) \end{aligned}$$

by Taylor expansions for v_0 and $\hat{G}$. Note that

$$\int_B \nabla_y w\, dy = \int_{\partial B} \frac{\partial w^-}{\partial \nu_y} y\, d\sigma_y$$

and

$$\psi = w + \nabla_x v_0(0) \cdot y = \frac{\partial v_0}{\partial x_l}(0) \phi_l,$$

where the ϕ_l are defined by (4). Hence

$$|B| \nabla_x v_0(0) + \int_B \nabla_y w\, dy = \int_B \nabla_y \psi\, dy,$$

from which we may rewrite (41) as

$$\begin{aligned} v_\alpha(x) - v_0(x) &= \alpha^3 \left(\frac{1}{\mu^1} - \frac{1}{\mu^0}\right) \nabla u_0(0) \cdot M\left(\frac{\mu^1}{\mu^0}\right) \nabla_z \hat{G}(x, 0) \\ &\quad + k^2 \left(1 - \frac{\varepsilon^1}{\varepsilon^0}\right) u_0(0) \hat{G}(x, 0) + o(\alpha^3) \qquad (42) \end{aligned}$$

for M defined by (4). By standard elliptic regularity, we obtain (32), where the term $o(\alpha^3)$ is uniform for $x \in \partial\Omega$.

We are now ready to prove Proposition 1. Integration by parts yields

$$\begin{aligned} &\int_{\partial\Omega} G(x, y) \frac{\partial}{\partial \nu_y} (\nabla_z \hat{G}(y, 0))\, d\sigma_y = \nabla_z G(x, 0) \text{ and} \\ &\int_{\partial\Omega} G(x, y) \frac{\partial}{\partial \nu_y} (\hat{G}(y, 0))\, d\sigma_y = G(x, 0). \qquad (43) \end{aligned}$$

By applying the operator S to (32) and using (43), we arrive at the promised asymptotic expansion (13), which, along with the boundedness of the operator S, implies that T_α converges to T_0 pointwise, which is the claim in point (a). Furthermore,

since the points z_j are away from the boundary $\partial\Omega$, it follows from (13) that the family of operators $T_\alpha - T_0$ is collectively compact, and so point (b) holds. Rewriting $T_\alpha = T_0 + (T_\alpha - T_0)$ and recalling that the operator T_0 is invertible, it follows immediately from [4] that T_α^{-1} is well defined, and point (c) in Proposition 1 holds.

Acknowledgment. The authors express their thanks to M. Vogelius for various beneficial discussions.

REFERENCES

[1] C. Alves and H. Ammari, *Boundary integral formulae for the reconstruction of imperfections of small diameter in an elastic medium*, SIAM J. Appl. Math., 62 (2001), pp. 94–106.

[2] H. Ammari, S. Moskow, and M. Vogelius, *Boundary integral formulas for the reconstruction of electromagnetic imperfections of small diameter*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 49–66.

[3] H. Ammari, M. Vogelius, and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of imperfections of small diameter* II. *The full Maxwell equations*, J. Math. Pures Appl. (9), 80 (2001), pp. 769–814.

[4] P. M. Anselone, *Collectively Compact Operator Approximation Theory and Applications to Integral Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1971.

[5] A. P. Calderon, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and Its Applications to Continuum Physics, Soc. Brasileira de Matemática, Rio de Janeiro, 1980, pp. 65–73.

[6] D. J. Cedio-Fengya, S. Moskow, and M. Vogelius, *Identification of conductivity imperfections of small diameter by boundary measurements. Continuous dependence and computional reconstruction*, Inverse Problems, 14 (1998), pp. 553–595.

[7] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Appl. Math. Sci. 93, Springer-Verlag, Berlin, 1992.

[8] G. Dassios, *Low-frequency moments in inverse scattering theory*, J. Math. Phys., 31 (1990), pp. 1691–1692.

[9] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.

[10] G. Eskin and J. Ralston, *The inverse backscattering in three dimensions*, Comm. Math. Phys., 124 (1989), pp. 169–215.

[11] G. Eskin and J. Ralston, *Inverse backscattering in two dimensions*, Comm. Math. Phys., 138 (1991), pp. 451–486.

[12] G. Eskin and J. Ralston, *Inverse backscattering*, J. Anal. Math., 58 (1992), pp. 177–190.

[13] A. Friedman and M. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Ration. Mech. Anal., 105 (1989), pp. 299–326.

[14] M. Ikehata, *Reconstruction of an obstacle from the scattering amplitude at a fixed frequency*, Inverse Problems, 14 (1998), pp. 949–954.

[15] V. Isakov, *Inverse Problems for Partial Differential Equations*, Appl. Math. Sci. 127, Springer-Verlag, New York, 1998.

[16] J. C. Nédélec, *Acoustic and Electromagnetic Equations. Integral Representations for Harmonic Problems*, Springer-Verlag, New York, 2001.

[17] R. G. Novikov, *Reconstruction of an exponentially decreasing potential for the three-dimensional Schrödinger equation through the scattering amplitude at a fixed energy*, C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 657–662.

[18] R. G. Novikov, *On determination of the Fourier transform of a potential from the scattering amplitude*, Inverse Problems, 17 (2001), pp. 1243–1251.

[19] P. Stefanov and G. Uhlmann, *Inverse backscattering for the acoustic equation*, SIAM J. Math. Anal., 28 (1997), pp. 1191–1204.

[20] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.

[21] J. Sylvester and G. Uhlmann, *The Dirichlet to Neumann map and applications*, in Inverse Problems in Partial Differential Equations (Arcata, CA, 1989), Proc. Appl. Math. 42, SIAM, Philadelphia, 1990, pp. 101–139.

[22] M. E. Taylor, *Partial Differential Equations* II. *Qualitative Studies of Linear Equations*, Appl. Math. Sci. 116, Springer-Verlag, New York, 1996.

[23] M. E. Taylor, *Estimates for approximate solutions to acoustic inverse scattering problems*, in Inverse Problems in Wave Propagation (Minneapolis, MN, 1995), IMA Vol. Math. Appl. 90, Springer-Verlag, New York, 1997, pp. 463–499.

[24] M. Vogelius and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 723–748.

[25] D. Volkov, *An Inverse Problem for the Time Harmonic Maxwell Equations*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 2001.

SIAM J. MATH. ANAL.
Vol. 34, No. 4, pp. 901–924

ON k-MONOTONE APPROXIMATION BY FREE KNOT SPLINES*

KIRILL KOPOTUN† AND ALEXEI SHADRIN‡

Abstract. Let $\mathbb{S}_{N,r}$ be the (nonlinear) space of free knot splines of degree $r-1$ with at most N pieces in $[a,b]$, and let $\mathcal{M}^k$ be the class of all k-monotone functions on (a,b), i.e., those functions f for which the kth divided difference $[x_0,\dots,x_k]f$ is nonnegative for all choices of $(k+1)$ distinct points $x_0,\dots,x_k$ in (a,b).

In this paper, we solve the problem of *shape preserving* approximation of k-monotone functions by splines from $\mathbb{S}_{N,r}$ in the $\mathbb{L}_p$-metric, i.e., by splines which are constrained to be k-monotone as well. Namely, we prove that the order of such approximation is essentially the same as that by the nonconstrained splines. Precisely, it is shown that, for every $k,r,N\in\mathbb{N}$, $r\ge k$, and any $0<p\le\infty$, there exist constants $c_0=c_0(r,k)$ and $c_1=c_1(r,k,p)$ such that

$$\operatorname{dist}(f,\mathbb{S}_{c_0N,r}\cap M^k)_p\le c_1\operatorname{dist}(f,\mathbb{S}_{N,r})_p\quad\forall f\in M^k.$$

This extends to all $k\in\mathbb{N}$ results obtained earlier by Leviatan and Shadrin and by Petrov for $k\le 3$.

Key words. free knot splines, constrained approximation, k-monotone approximation, approximation order, Markov moment problem, Whitney-type estimates

AMS subject classifications. 41A15, 41A25, 41A29, 41A05, 65D05, 65D07

PII. S0036141002358514

1. Introduction and main results.

In this paper, we solve the problem of *shape preserving* approximation of k-monotone functions by splines with free knots in the $\mathbb{L}_p$-metric, i.e., by splines which are constrained to be k-monotone as well. Namely, we prove that the order of such approximation is essentially the same as that for the nonconstrained splines, thus confirming expectations of some standing.

Given $k\in\mathbb{Z}_+$ and an interval $I=(a,b)$, a function $f:I\mapsto\mathbb{R}$ is said to be k-monotone on I if its kth divided differences $[x_0,\dots,x_k]f$ are nonnegative for all choices of $(k+1)$ distinct points $x_0,\dots,x_k$ in I. We denote the class of all such functions by $\mathcal{M}^k:=\mathcal{M}^k(I)$. Thus $f\in\mathcal{M}^0$ is nonnegative, $f\in\mathcal{M}^1$ is nondecreasing, and $f\in\mathcal{M}^2$ is a convex function. If $f\in\mathbb{C}^k(I)$, then $f\in\mathcal{M}^k$ if and only if $f^{(k)}\ge 0$ on I.

We would like to emphasize that functions from $\mathcal{M}^k$ are not assumed to be defined at the endpoints of the interval (a,b) and hence have to be neither bounded nor integrable on (a,b). For example, if $f(x)=(-1)^kx^{-1-1/p}$, then $f\in\mathcal{M}^k(0,1)$ for $k\in\mathbb{N}$, but $f\notin\mathbb{L}_p(0,1)$, $0<p\le\infty$. (Throughout the paper, $\mathbb{L}_\infty(I)$ denotes the space of all measurable essentially bounded functions equipped with the norm $\|f\|_{\mathbb{L}_\infty(I)}:=\operatorname{ess\,sup}_I|f|$.)

Hence we now define $\mathcal{M}^k_p:=\mathcal{M}^k\cap\mathbb{L}_p$ and also remark that the functions from the cone $\mathcal{M}^k$ are sometimes referred to as "k-convex."

Let $f\in\mathcal{M}^k_p$ and $\mathcal{U}$ be a subset of $\mathbb{L}_p$. The best (nonconstrained) approximation of f from $\mathcal{U}$ is defined by

$$E(f,\mathcal{U})_p:=\inf_{u\in\mathcal{U}}\|f-u\|_p.$$

*Received by the editors February 18, 2002; accepted for publication (in revised form) August 30, 2002; published electronically March 5, 2003.
http://www.siam.org/journals/sima/34-4/35851.html

†Department of Mathematics, University of Manitoba, Winnipeg, Manitoba, R3T 2N2, Canada (kopotunk@cc.umanitoba.ca). The work of this author was supported by NSERC of Canada.

‡Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, CB3 9EW, UK (a.shadrin@damtp.cam.ac.uk).

In contrast, in k-monotone approximation, one is interested in the value

$$E^{(k)}(f,\mathcal{U})_p := \inf_{u\in\mathcal{U}\cap\mathcal{M}^k} \|f-u\|_p.$$

That is, approximants are assumed to preserve the k-monotone shape of f. Clearly, the shape preserving approximation is more restrictive; hence $E^{(k)}(f,\mathcal{U})_p \ge E(f,\mathcal{U})_p$ for all $f\in\mathcal{M}^k$ and $\mathcal{U}\subset\mathbb{L}_p$. Is it much worse? Lorentz and Zeller [10], [11] proved that, for $\mathcal{U}=\Pi_n$, the space of all algebraic polynomials of order n, any $k\in\mathbb{N}$, and any constant $c_0\in\mathbb{N}$, there exists a function $f\in\mathcal{M}^k_p$ such that

$$\frac{E^{(k)}(f,\Pi_{c_0 n})_p}{E(f,\Pi_n)_p}\to\infty \quad \text{as} \quad n\to\infty. \tag{1.1}$$

(The same estimate is true for a sequence of any reasonable linear subspaces $\mathcal{U}_n$ instead of Π_n.) On the other hand, monotone and convex polynomial approximations allow Jackson-type estimates, for example,

$$E^{(k)}(f,\Pi_n)_\infty \le c_k\omega_{k+1}(f,\tfrac{1}{n})_\infty, \quad k=1,2,$$

but they have essential restrictions (as well as gaps) in comparison with the nonconstrained estimates.

Splines with free knots, $s\in\mathcal{S}_{N,r}$, are piecewise polynomials of order r (degree $r-1$), where only the number of pieces, N at most, not their position, is prescribed. (Note that we do not make any assumptions about the smoothness of functions in $\mathcal{S}_{N,r}$.) They are a classical tool of *nonlinear* approximation (along with the rational functions). As that, they achieve a better rate of approximation compared with the linear methods. The simplest example (see, e.g., [4, p. 365]) is that

$$E(f,\mathcal{S}_{N,1})_\infty \le \frac{K}{2N} \quad\Leftrightarrow\quad \operatorname{Var}_{[0,1]}(f)\le K,$$

whereas, for $\mathbb{L}_\infty$-approximation by piecewise constants with N *equidistant* knots, the rate $\mathcal{O}(N^{-1})$ is attained only for $\mathbb{W}^1_\infty$, roughly the class of continuously differentiable functions, which is much narrower than the class of functions of bounded variation.

It was DeVore who had much advocated the studies of the nonlinear methods in k-monotone approximation. Set

$$E_{N,r}(f)_p := E(f,\mathcal{S}_{N,r})_p, \quad E^{(k)}_{N,r}(f)_p := E^{(k)}(f,\mathcal{S}_{N,r})_p.$$

Notice that, since $\mathcal{M}^k(0,1)\subset\mathbb{C}^{k-2}(0,1)$ (see Lemma 3.1), the set $\mathcal{S}_{N,r}\cap\mathcal{M}^k$ contains functions other than k-monotone polynomials of order r only if $r\ge k$. In 1995, Leviatan and Shadrin [9] and Petrov [14] independently proved that, for $k=1,2$, $r\ge k$, and $0<p\le\infty$, there exists a constant $c_0=\mathcal{O}(r)$ such that, for any $f\in\mathcal{M}^k_p$, $k=1,2$,

$$E^{(k)}_{c_0 N,r}(f)_p \le E_{N,r}(f)_p. \tag{1.2}$$

This result showed that the order of monotone and convex approximation by free knot splines is essentially the same as that in the nonconstrained case, which, in view of (1.1), is a striking contrast to the linear approximation methods. Naturally, one would expect that the situation is similar for $k\ge 3$. However, the technique used in [9], [14] was based on some explicit constructions and some properties of monotone

and convex functions which have no straightforward analogues for general k. (Say, for $k = 1, 2$, the maximum of two k-monotone functions is a k-monotone function, while this is no longer true for larger k.) Petrov [15] has managed to adopt this technique for $k = 3$ and $p = \infty$, obtaining an analogue of (1.2), but it became clear that, for general $k \in \mathbb{N}$, new ideas are required.

Here, we prove the following general result.

THEOREM 1.1. *Let $k, r, N \in \mathbb{N}$, $r \geq k$, and $0 < p \leq \infty$. Then there exist constants $c_0 \leq C(k)\max(1, r-k)$ and $c_1 = c_1(r, k, p)$ such that, for all $f \in \mathcal{M}_p^k$,*

$$E^{(k)}_{c_0 N, r}(f)_p \leq c_1 E_{N,r}(f)_p. \tag{1.3}$$

Using [9, Lemma 3], the following result on k-monotone approximation by smooth splines is an immediate corollary of Theorem 1.1.

COROLLARY 1.2. *Let $k, r, N \in \mathbb{N}$, $r \geq k$, and $0 < p \leq \infty$, and denote $\widetilde{E}^{(k)}_{N,r}(f)_p := E^{(k)}(f, \mathcal{S}_{N,r} \cap \mathbb{C}^{(r-2)})_p$. Then there exist constants $c_0 \leq C(k)\max(1, r-k)$ and $c_1 = c_1(r, k, p)$ such that, for all $f \in \mathcal{M}_p^k$,*

$$\widetilde{E}^{(k)}_{c_0 N, r}(f)_p \leq c_1 E_{N,r}(f)_p.$$

For $k = 1$ and 2, Theorem 1.1 is an immediate consequence of (1.2). Because functions in $\mathcal{M}^1(a, b)$ (unlike those in $\mathcal{M}^k(a, b)$ with $k \geq 2$) do not have to be continuous everywhere on (a, b), the case $k = 1$ is somewhat different from $k \geq 2$ (though constructions are much simpler, and some auxiliary statements become trivial if one lets k be equal to 1). Thus, in order to make this paper more readable, we concentrate below only on the more difficult case for $k \geq 2$. At the same time, we mention that some of the statements are valid or can be modified to become valid for $k = 1$ as well.

Now, all direct results for the best (unconstrained) free knot spline approximation are being readily extended for the k-monotone case.

COROLLARY 1.3. *Let $k, r, N \in \mathbb{N}$, $r \geq k$, and let $f \in \mathcal{M}_\infty^k$ be such that $f^{(r-1)}$ is of bounded variation on $[0, 1]$. Then*

$$E^{(k)}_{N,r}(f)_\infty \leq c(r, k) N^{-r} \operatorname{Var}_{[0,1]}(f^{(r-1)}).$$

This corollary is an immediate consequence of Theorem 1.1 and [4, Theorem 12.4.5]. It is related to an earlier result of Hu [5] which was actually the first result in k-monotone approximation by free knot splines: *For $f \in \mathbb{W}_1^r \cap \mathcal{M}_\infty^k$, the order of k-monotone approximation by $S_{N,r}$ in $\mathbb{L}_\infty$ is $\mathcal{O}(N^{-r})$.*

The following corollary follows from Petrushev's estimate of (unconstrained) free knot spline approximation (see [16], [17, Theorem 7.3], and [4, Theorem 12.8.2]).

COROLLARY 1.4. *Let $k, r, N \in \mathbb{N}$, $r \geq k$, $0 < p < \infty$, and $0 < \alpha < r$. Then, if $f \in \mathcal{M}_p^k \cap B^\alpha$,*

$$E^{(k)}_{N,r}(f)_p \leq c(\alpha, p, r) N^{-\alpha} |f|_{B^\alpha},$$

where $B^\alpha := B^\alpha_\gamma(L_\gamma)$, $1/\gamma = \alpha + 1/p$, denotes the Besov space with the seminorm $|f|_{B^\alpha}$ defined by

$$|f|_{B^\alpha} = \left(\int_0^\infty t^{-\alpha\gamma - 1} \omega_r(f, t)_\gamma^\gamma \, dt \right)^{1/\gamma}.$$

Let us comment on the constants c_0, c_1 involved in (1.3), namely, on the question of whether it is possible to have any (or both) of them equal to 1.

Leviatan and Shadrin [9] showed that, in order to retain the same degree of approximation for the k-monotone free knot splines approximation as for the best one, the increase of the knot number is unavoidable if $r \geq k+2$. Precisely, *for any* $r \geq k+2$, $N \in \mathbb{N}$, $0 < p \leq \infty$, *any* $c > 0$, *and* $c_* = 2\lfloor \frac{r-k}{2} \rfloor$, *there exists a function* $f \in \mathcal{M}^k_\infty$ *such that*

$$E^{(k)}_{c_* N,r}(f)_p > cE_{N,r}(f)_p, \quad r \geq k+2.$$

Thus the question about whether or not it is necessary to increase the number of knots remains open only for $r = k$ and $k+1$.

On the other hand, for $r = k$ and $p = \infty$, a part of a theorem by Johnson (see Braess [2, Theorem VIII.3.4, p. 238]) is that *for any* k, *the best free knot spline approximant of order* k *to a* k*-monotone function in the* $\mathbb{L}_\infty$*-norm is* k*-monotone itself*; i.e., in this case, $c_0 = c_1 = 1$, $r = k$, and, for any $f \in \mathcal{M}^k_\infty$,

$$E^{(k)}_{N,k}(f)_\infty = E_{N,k}(f)_\infty. \tag{1.4}$$

It would be interesting to find the exact order of $c_0(r,k)$ as a function of r and k. Estimates (1.2) and (1.4) also suggest another question—namely, whether the value $c_1 = 1$ in (1.3) can be attained with some $c_0' = c_0'(r,k)$.

Notation. We let $I = (a,b)$ if not stated otherwise and set $\mathbb{L}_p := \mathbb{L}_p(I)$, $\|\cdot\|_p := \|\cdot\|_{\mathbb{L}_p(I)}$, $\mathcal{S}_{N,k} := \mathcal{S}_{N,k}(I)$, etc.; i.e., the interval I is omitted if there is no risk of confusion.

Further, $f^{(i)}(x+)$ and $f^{(i)}(x-)$ denote the right and the left ith derivatives of f at x, respectively.

$c_{p,r,k}$ and $c(p,r,k)$ stand for a constant which depends only on the parameters given (p, r, and k in this case), where, for $0 < p \leq \infty$, dependence on p means dependence on $\min(1,p)$.

The "prime" notation k' is going to be reserved for $\lfloor k/2 \rfloor + 1$ throughout this paper:

$$k' := \lfloor k/2 \rfloor + 1.$$

For $f \in \mathbb{L}_p(a,b)$ and a set $\mathcal{U} \subset \mathbb{L}_p(a,b)$, we define

$$\mathcal{P}_\mathcal{U}(f)_p := \mathcal{P}_\mathcal{U}(f)_{\mathbb{L}_p(a,b)} := \{u \in \mathcal{U} : \ \|f-u\|_p = E(f,\mathcal{U})_p\}.$$

In other words, $\mathcal{P}_\mathcal{U}(f)_p$ is the set of all best $\mathbb{L}_p$-approximants to f from $\mathcal{U}$ on (a,b).

2. Outline of the proof. The general direction of the proof is the same as it was for $k = 1, 2$: given a k-monotone function f, one takes $\sigma \in \mathcal{P}_{S_{N,r}}(f)_p$, a best free knot spline approximant to f (which is not necessarily k-monotone), and puts some corrections into it trying to convert it into a k-monotone spline preserving the approximation order. For $k = 1$ and 2, these corrections were done by explicit constructions which, unfortunately, have no straightforward generalizations for $k \geq 3$, and so our basic idea came from the following general considerations.

There is another notion of k-monotone approximation in which a function f which is not in $\mathcal{M}^k$ is approximated by elements from the entire $\mathcal{M}^k$. ($\mathcal{M}^k$ is a convex cone.) There is an extended literature on this subject, where one studies existence

and uniqueness of best k-monotone approximant of this type, its characterization, and its structural properties; see, e.g., [19] and the references therein. When can one have a need to approximate an arbitrary function by a k-monotone one? The only situation we can think of is the necessity to correct the data which must be k-monotone by some a priori assumptions. This is exactly the case of shape preserving approximation, and this is how we correct σ.

Given $f \in \mathcal{M}^k$, we take $\sigma \in \mathcal{P}_{\mathcal{S}_{N,r}}(f)_p$, a best free knot spline approximant to f, and correct σ by $\mathfrak{f}_* \in \mathcal{P}_{\mathcal{M}^k}(\sigma)_p$, a best approximant to σ from $\mathcal{M}^k$.

Here are two observations concerning this idea.

(1) *Approximation property of* $\mathfrak{f}_*$. The function f belongs to $\mathcal{M}^k$, but $\mathfrak{f}_*$ is a best approximant to σ from $\mathcal{M}^k$; hence

$$\|\sigma - \mathfrak{f}_*\|_p \le \|\sigma - f\|_p.$$

Therefore,

$$c_p\|f - \mathfrak{f}_*\|_p \le \|f - \sigma\|_p + \|\sigma - \mathfrak{f}_*\|_p \le 2\|f - \sigma\|_p;$$

i.e., $\mathfrak{f}_*$ approximates f as well as σ.

(2) *Spline structure of* $\mathfrak{f}_*$. A result from the theory of approximation by elements of $\mathcal{M}^k$ reads that (in the "piecewise sense") either $\mathfrak{f}_*$ is identical with σ (which is a spline of order r) or it is a spline of order k (because the functions $g(x) = \sum_\alpha c_\alpha(x - x_\alpha)_+^{k-1}$, $c_\alpha > 0$, are the boundary points of the cone $\mathcal{M}^k$). Thus $\mathfrak{f}_*$ is a spline of order r. If $\mathfrak{f}_*$ had $\mathcal{O}(N)$ knots, then we could stop at this point. The problem is that it may have too many knots (infinitely many, in fact).

The paper is organized as follows.

(1) First, to ease the exposition, we switch to a local version of the idea described above and correct separately each polynomial part of σ by its best approximation $\mathfrak{f}_*$ from $\mathcal{M}^k[f]$, a subclass of k-monotone functions defined locally (see section 3 for precise definition of $\mathcal{M}^k[f]$).

(2) In section 3, we cite some known results concerning existence and structure of the elements $\mathfrak{f}_* \in \mathcal{P}_{\mathcal{M}^k[f]}(\sigma)_p$. As mentioned earlier, $\mathfrak{f}_*$ is a spline of order r, but it may have too many knots to be in $\mathcal{S}_{cN,r}$, in which case we modify it into an appropriate spline s.

(3) Properties of s are formulated as Proposition 4.2 in section 4, where we use them to prove Theorem 1.1.

(4) The proof of Proposition 4.2 takes the rest of the paper. In sections 5–7, we blend $\mathfrak{f}_*$ with the polynomial parts of σ using some results from the theory of moments and consider some general aspects of this procedure. In section 8, we prepare to show that the blending spline s approximates f as well as $\mathfrak{f}_*$, and the final section 9 joins all the parts of the proof together.

Remark 2.1. The number of knots of $\mathfrak{f}_* \in \mathcal{P}_{\mathcal{M}^k[f]}(\sigma)_p$ is approximately the same as the number of distinct zeros of $\sigma - \mathfrak{f}_*$ (see Lemma 3.8). In our proofs, we assume that this number may be arbitrarily large. However, we conjecture that this is not the case; i.e., a best k-monotone approximant to a piecewise polynomial σ (and perhaps to any piecewise k-monotone function) with M pieces has only $\mathcal{O}(M)$ points of intersection with σ. If this is indeed the case, then there is no need for the considerations given in sections 5–9. This conjecture is true for $k = 1, 2$ as one can easily check, and our method gives a simpler proof for these cases than in [9] and [14]. For $k \ge 3$, the problem is open.

Remark 2.2. Actually, the correction of σ made explicitly for $k = 1, 2$ in [9] and [14] is exactly the best k-monotone approximation of σ from $\mathcal{M}^k[f]$ under the additional restriction that this is also a one-sided approximation. This restriction provides the constant $c_1 = 1$ on the right-hand side of (1.2). For $k \geq 3$, we cannot pose such a restriction; hence $c_1 > 1$ in (1.3).

3. Classes $\mathcal{M}^k[f]$ and their properties. The following lemma lists some basic properties of k-monotone functions for $k \geq 2$.

LEMMA 3.1. *The following statements are equivalent for* $k \geq 2$:

(0) $f \in \mathcal{M}^k(0,1)$.

(1) $f^{(k-2)}$ *exists and is convex on* $(0,1)$.

(2) $f^{(k-2)}$ *is absolutely continuous on any closed subinterval of* $(0,1)$ *and has left and right derivatives,* $f^{(k-1)}(\cdot-)$ *and* $f^{(k-1)}(\cdot+)$*, which are, respectively, left- and right-continuous and nondecreasing on* $(0,1)$.

(3) *For each closed subinterval* $[a,b] \subset (0,1)$*, there is a polynomial* $p \in \Pi_k$ *and a bounded nondecreasing function* μ *such that*

$$f(x) = p(x) + \frac{1}{k!}\int_a^b k(x-t)_+^{k-1}\,d\mu(t), \quad x \in [a,b].$$

Proof. See Bullen [3, Theorem 7, Corollary 8]. See also [13], [18] for various properties of k-monotone functions (called "k-convex" there) and their applications. ∎

Lemma 3.1(2) allows us to introduce the following classes of function.

By $\mathcal{M}^k_{a+} := \mathcal{M}^k_{a+}(a,b)$ and $\mathcal{M}^k_{b-} := \mathcal{M}^k_{b-}(a,b)$ we denote the subclasses of those functions $f \in \mathcal{M}^k(a,b)$ for which the values $\{f^{(i)}(a+)\}_{i=0}^{k-1}$ and $\{f^{(i)}(b-)\}_{i=0}^{k-1}$, respectively, are finite, and we set $\mathcal{M}^k_* := \mathcal{M}^k_*(a,b) := \mathcal{M}^k_{a+} \cap \mathcal{M}^k_{b-}$.

For $f \in \mathcal{M}^k_{a+}$ and $g \in \mathcal{M}^k_{b-}$, we define

$$\mathcal{M}^k_{a+}[f] := \{h \in \mathcal{M}^k \mid\ h^{(i)}(a+) = f^{(i)}(a+),\ i = 0,\dots,k-2;\ h^{(k-1)}(a+) \geq f^{(k-1)}(a+)\},$$
$$\mathcal{M}^k_{b-}[g] := \{h \in \mathcal{M}^k \mid\ h^{(i)}(b-) = g^{(i)}(b-),\ i = 0,\dots,k-2;\ h^{(k-1)}(b-) \leq g^{(k-1)}(b-)\}.$$

Finally, let

$$\mathcal{M}^k[f,g] = \mathcal{M}^k_{a+}[f] \cap \mathcal{M}^k_{b-}[g],$$

and, for $f \in \mathcal{M}^k_*$,

$$\mathcal{M}^k[f] = \mathcal{M}^k[f,f].$$

Note that $\mathcal{M}^k[f]$ is always nonempty (it contains f), while $\mathcal{M}^k[f,g]$ can be the empty set. In section 7, we give a sufficient condition on f and g which guarantees that there is a function h from $\mathcal{M}^k[f,g]$.

LEMMA 3.2. *Let* $f, g \in \mathcal{M}^k(0,1)$*, and let* $[a,b] \subset (0,1)$*. Then* $f, g \in \mathcal{M}^k_*(a,b)$*, and, for any* $h \in \mathcal{M}^k[f,g](a,b)$ *(if it exists), the function*

$$\widetilde{h}(x) := \begin{cases} f(x), & x \in (0,a], \\ h(x), & x \in (a,b), \\ g(x), & x \in [b,1), \end{cases}$$

belongs to $\mathcal{M}^k(0,1)$.

Proof. The proof is an immediate consequence of Lemma 3.1(2). ∎

We will use Lemma 3.2 without further reference to build k-monotone functions from k-monotone pieces. For example, if $f \in \mathcal{M}^k_*(0,1)$, $\cup I_\ell = (0,1)$ with $I_\ell \cap I_{\ell'} = \emptyset$ if $\ell \neq \ell'$, and $h_\ell \in \mathcal{M}^k[f](I_\ell)$, then the function h, defined as $h := h_\ell$ on I_ℓ, belongs to $\mathcal{M}^k[f](0,1)$.

Now we consider some properties of approximation from $\mathcal{M}^k[f]$.

LEMMA 3.3. *Let $k \geq 2$, $0 < p \leq \infty$, and $f \in \mathcal{M}^k_*(a,b)$. Then, for any $g \in \mathbb{L}_p$, an element of its best $\mathbb{L}_p$-approximation from $\mathcal{M}^k[f]$ exists; i.e., the set $\mathcal{P}_{\mathcal{M}^k[f]}(g)_p$ is not empty.*

Proof. The proof is based on arguments similar to those used by Zwick [20, Theorem 4] for the case $p = \infty$. We give it here for completeness. Set

$$\alpha_i := f^{(i)}(a+) \quad \text{and} \quad \beta_i := f^{(i)}(b-), \quad i = 0, \dots, k-1,$$

and consider a sequence $(f_j) \subset \mathcal{M}^k[f]$ such that, for $j \in \mathbb{N}$,

$$\|f_j - g\|_p^p \leq E(g, \mathcal{M}^k[f])_p^p + 1/j \quad \text{if} \quad 0 < p < 1$$

and

$$\|f_j - g\|_p \leq E(g, \mathcal{M}^k[f])_p + 1/j \quad \text{if} \quad 1 \leq p \leq \infty.$$

Since $f_j^{(k-2)}(x) = \alpha_{k-2} + \int_a^x f_j^{(k-1)}(t)\,dt$ and $\|f_j^{(k-1)}\|_\infty \leq \max\{|\alpha_{k-1}|, |\beta_{k-1}|\}$, we conclude that $(f_j^{(k-2)})$ is uniformly bounded and equicontinuous on $[a,b]$. Therefore, there exists a subsequence $(f_{j_s}^{(k-2)})$ which converges to a function h_* uniformly on $[a,b]$, and this h_* is necessarily convex and satisfies $h_*'(a+) \geq \alpha_{k-1}$ and $h_*'(b-) \leq \beta_{k-1}$. Now, the function $\mathfrak{f}_*$ such that $\mathfrak{f}_* := h_*$, if $k = 2$, and

$$\mathfrak{f}_*(x) := \sum_{i=0}^{k-3} \frac{\alpha_i}{i!}(x-a)^i + \frac{1}{(k-3)!}\int_a^b (x-t)_+^{k-3} h_*(t)\,dt, \quad k \geq 3,$$

is in $\mathcal{M}^k[f]$ and satisfies $\|g - \mathfrak{f}_*\|_p = E(g, \mathcal{M}^k[f])_p$, i.e., $\mathfrak{f}_* \in \mathcal{P}_{\mathcal{M}^k[f]}(g)_p$. ∎

LEMMA 3.4 (Zwick [21]). *Let $k \in \mathbb{N}$ and $f \in \mathcal{M}^k_*(a,b)$. Then there exist two splines $z_\nu = z_\nu(f,[a,b])$, $\nu = 1,2$, such that*

$$z_1, z_2 \in \mathcal{M}^k[f] \cap \mathcal{S}_{k',k}, \quad k' = \lfloor k/2 \rfloor + 1,$$

and

$$z_1 \leq f \leq z_2 \quad \text{on} \quad [a,b].$$

If f does not belong to $\mathcal{S}_{k',k}$, then the inequalities are strict, respectively, on some nonempty intervals I_1, I_2 in $[a,b]$.

Remark 3.5. In [21], more precise conclusions regarding the number of polynomial pieces k' of the splines z_ν and their boundary values are given. The proof is based on the Markov–Krein theorem from the theory of moments.

Remark 3.6. We emphasize that k' denotes $\lfloor k/2 \rfloor + 1$ throughout this paper.

A simple, yet important, consequence of Lemma 3.4 is the following result on the structural properties of best $\mathbb{L}_p$-approximants from $\mathcal{M}^k[f]$.

LEMMA 3.7. *For $k \geq 2$, $0 < p < \infty$, and $I = (a,b)$, let $g \in \mathbb{C}$, $f \in \mathcal{M}^k_*$, and $\mathfrak{f}_* \in \mathcal{P}_{\mathcal{M}^k[f]}(g)_p$. If the difference $g - \mathfrak{f}_*$ has no zeros inside an interval $(c,d) \subset (a,b)$, then $\mathfrak{f}_* \in \mathcal{S}_{k',k}[c,d]$.*

Proof. The idea of the proof is similar to what was considered by Zwick [21] in the case for $p = 1$. Suppose that $0 < p < \infty$. Without loss of generality, we can assume that $\mathfrak{f}_*(x) > g(x)$, $x \in (c, d)$. Now suppose that $\mathfrak{f}_* \notin \mathcal{S}_{k',k}[c+\epsilon, d-\epsilon]$ for some $\epsilon > 0$. Consider a function $\tilde{f}$ obtained from $\mathfrak{f}_*$ by replacing it on the interval $[c+\epsilon, d-\epsilon]$ by $z_1(\mathfrak{f}_*, [c+\epsilon, d-\epsilon])$ (see Lemma 3.4). Then $\tilde{f} \in \mathcal{M}^k[f]$ and $\mathfrak{f}_* - \tilde{f} \geq 0$ on $[a, b]$ with this inequality being strict on a nonempty interval contained in $(c+\epsilon, d-\epsilon)$.

Since $\mathfrak{f}_* - g$ is a continuous positive function on a closed interval $[c+\epsilon, d-\epsilon]$, there exists $\delta > 0$ such that $\mathfrak{f}_*(x) \geq g(x) + \delta$, $x \in [c+\epsilon, d-\epsilon]$. Therefore, there exists $0 < \mu < 1$ such that $\widehat{f}(x) := \mu\mathfrak{f}_*(x) + (1-\mu)\tilde{f}(x)$ satisfies the inequalities $g(x) < \widehat{f} \leq \mathfrak{f}_*$ on $[c+\epsilon, d-\epsilon]$, and $\|\mathfrak{f}_* - \widehat{f}\|_{\mathbb{L}_p[c+\epsilon, d-\epsilon]} \neq 0$. This implies that $\|\widehat{f} - g\|_p < \|\mathfrak{f}_* - g\|_p$, which contradicts our assumption that $\mathfrak{f}_* \in \mathcal{P}_{\mathcal{M}^k[f]}(g)_p$.

Hence $\mathfrak{f}_* \in \mathcal{S}_{k',k}[c+\epsilon, d-\epsilon]$ for all $\epsilon > 0$, which implies that $\mathfrak{f}_* \in \mathcal{S}_{k',k}[c, d]$. ∎

LEMMA 3.8. *For $k \geq 2$, $0 < p < \infty$, and $I = (a, b)$, let $g \in \mathbb{C}$, $f \in \mathcal{M}^k_*$, and $\mathfrak{f}_* \in \mathcal{P}_{\mathcal{M}^k[f]}(g)_p$. Further, let $\mathfrak{Z}$ be the set of zeros of $g - \mathfrak{f}_*$, i.e.,*

$$\mathfrak{Z} := \{z \in I \mid \ g(z) = \mathfrak{f}_*(z)\},$$

and let $\mathfrak{Z}^*$ be the set of all limit points of $\mathfrak{Z}$. Then the following are true.

(1) *$\mathfrak{f}_* = g$ on $\mathfrak{Z}^*$.*

(2) *If, for a closed interval $[c, d] \subset I \setminus \mathfrak{Z}^*$, the difference $g - \mathfrak{f}_*$ has (necessarily finitely many) $m - 1$ distinct zeros in (c, d), then $\mathfrak{f}_* \in \mathcal{S}_{mk',k}[c, d]$.*

Proof. This lemma is a variation of Zwick [21, Theorem 2]. In a similar form (though with $\mathfrak{Z}^*$ defined differently), it appeared in Damas and Marano [12]. Part 1 immediately follows from the continuity of g and $\mathfrak{f}_*$. Part 2 is a consequence of Lemma 3.7. ∎

For $p = \infty$, Lemma 3.7 is not valid because local changes influence the integral's value but not necessarily the sup-norm; hence there may be best k-monotone $\mathbb{L}_\infty$-approximants with a structure different from that specified in Lemma 3.8. However, for our purposes, it is enough that there is at least one element from $\mathcal{P}_{\mathcal{M}^k[f]}(g)_\infty$ that has the spline structure. The following statement is valid.

LEMMA 3.9. *For $k \geq 2$, $p = \infty$, and $I = (a, b)$, let $g \in \mathbb{C}$ and $f \in \mathcal{M}^k_*$. Then there exists $\mathfrak{f}_* \in \mathcal{P}_{\mathcal{M}^k[f]}(g)_\infty$ such that all the conclusions of Lemma 3.8 hold true.*

Proof. The idea of the proof is to take as $\mathfrak{f}_*$ an element which minimizes, say, the L_2-norm of $g - \mathfrak{f}_*$ over $\mathfrak{f}_* \in \mathcal{P}_{\mathcal{M}^k[f]}(g)_\infty$. We omit the details. ∎

Now the spline structure of the best k-monotone approximant to any spline readily follows.

COROLLARY 3.10. *For $r \geq k \geq 2$, $0 < p \leq \infty$, and $I = (a, b)$, let $g \in \mathcal{S}_{N,r} \cap \mathbb{C}$ and $f \in \mathcal{M}^k_*$. Then there is an $\mathfrak{f}_* \in \mathcal{P}_{\mathcal{M}^k[f]}(g)_p$ which is a piecewise polynomial of order r.*

4. Proof of Theorem 1.1. The following three propositions are the main components of the proof.

PROPOSITION 4.1. *For $k, r \in \mathbb{N}$, $r \geq k \geq 2$, $0 < p \leq \infty$, and $I = (a, b)$, let $f \in \mathcal{M}^k_*$ and $(-\mathfrak{p}) \in (\Pi_r \setminus \Pi_k) \cap \mathcal{M}^k$. Then there exists a spline s such that*

$$s \in \mathcal{S}_{(k+1)k',k} \cap \mathcal{M}^k[f]$$

and

$$\|\mathfrak{p} - s\|_p = E(\mathfrak{p}, \mathcal{M}^k[f])_p.$$

Proof. Let us show that $\mathfrak{f}_*$, a best approximant to $\mathfrak{p}$ from $\mathcal{M}^k[f]$, satisfies all the conclusions of the proposition (hence $s := \mathfrak{f}_*$). Since, by the definition,

$$\mathfrak{f}_* \in \mathcal{M}^k[f], \quad \|\mathfrak{p} - \mathfrak{f}_*\|_p = E(\mathfrak{p}, \mathcal{M}^k[f])_p,$$

only the spline structure needs to be proved. Since $(-\mathfrak{p})$ is a k-monotone polynomial of degree $> k-1$, it is a strictly k-monotone function in the sense that $(-\mathfrak{p})^{(k-2)}$ is strictly convex. Hence the function $(\mathfrak{f}_* - \mathfrak{p})^{(k-2)}$ is strictly convex too; thus it has at most two zeros, and, therefore, $\mathfrak{f}_* - \mathfrak{p}$ has not more than k distinct zeros on I. By Lemma 3.8 (or Lemma 3.9 in the case for $p = \infty$), $\mathfrak{f}_* \in \mathcal{S}_{(k+1)k',k}$, and the proof is complete. □

PROPOSITION 4.2. *Let $k, r \in \mathbb{N}$, $r \geq k \geq 2$, $0 < p \leq \infty$, $I = (a, b)$, $f \in \mathcal{M}^k_*$, and $\mathfrak{p} \in \Pi_r \cap \mathcal{M}^k$. Then there exist a constant $C(k)$ independent of I and a spline $s \in \mathcal{S}_{C(k),r} \cap \mathcal{M}^k[f]$ such that*

$$\|\mathfrak{p} - s\|_p \leq c_2 E(\mathfrak{p}, \mathcal{M}^k[f])_p, \quad c_2 = c_2(p, r, k).$$

Now, $\mathfrak{f}_*$ from $\mathcal{P}_{\mathcal{M}^k[f]}(\mathfrak{p})_p$ is still a piecewise polynomial of order r, but we cannot take $s = \mathfrak{f}_*$ because two k-monotone functions ($\mathfrak{f}_*$ and $\mathfrak{p}$ in our case) may have any number of intersections; hence $\mathfrak{f}_*$ may have any number of knots. We obtain s as a modification of $\mathfrak{f}_*$, which will be done in the following sections with the proof of Proposition 4.2 given in section 9.

PROPOSITION 4.3. *Let $k, r \in \mathbb{N}$, $r \geq k \geq 2$, $0 < p \leq \infty$, $I = (a, b)$, and $f \in \mathcal{M}^k_*$, and let $\mathfrak{p}$ be such that either $\mathfrak{p} \in \Pi_r \cap \mathcal{M}^k$ or $(-\mathfrak{p}) \in (\Pi_r \setminus \Pi_k) \cap \mathcal{M}^k$. Then there exists a spline s such that*

$$s \in \mathcal{S}_{C(k),r} \cap \mathcal{M}^k[f]$$

and

$$\|f - s\|_p \leq c_1 \|f - \mathfrak{p}\|_p, \quad c_1 = c_1(p, r, k). \tag{4.1}$$

Proof. Let s be the spline from either of Propositions 4.1 and 4.2 so that $s \in \mathcal{S}_{C(k),r} \cap \mathcal{M}^k[f]$ and

$$\|\mathfrak{p} - s\|_p \leq c_2 E(\mathfrak{p}, \mathcal{M}^k[f])_p. \tag{4.2}$$

We need only to prove (4.1). Using the triangle inequality and the estimate (4.2), we obtain

$$c_p \|f - s\|_p \leq \|f - \mathfrak{p}\|_p + \|\mathfrak{p} - s\|_p \leq \|f - \mathfrak{p}\|_p + c_2 E(\mathfrak{p}, \mathcal{M}^k[f])_p.$$

Since f belongs to $\mathcal{M}^k[f]$ in a trivial manner, it follows that

$$E(\mathfrak{p}, \mathcal{M}^k[f])_p := \inf_{u \in \mathcal{M}^k[f]} \|\mathfrak{p} - u\|_p \leq \|\mathfrak{p} - f\|_p.$$

Thus

$$c_p \|f - s\|_p \leq (c_2 + 1) \|f - \mathfrak{p}\|_p. \qquad \square$$

Finally, the following lemma shows that, in the proof of Theorem 1.1, instead of an arbitrary $f \in \mathcal{M}^k_p(0, 1)$, we may consider $f \in \mathcal{M}^k_*(0, 1)$; i.e., we may assume that the function f and its derivatives are bounded at the endpoints.

LEMMA 4.4. *Let $k \in \mathbb{N}$, $0 < p \le \infty$, and $f \in \mathcal{M}_p^k(0,1)$. Then, for any $\epsilon > 0$, there exists $f_\epsilon \in \mathcal{M}_*^k(0,1)$ such that*

$$\|f - f_\epsilon\|_p < \epsilon.$$

Proof. For $f \in \mathcal{M}_p^k(0,1)$ and $x_0 \in (0,1)$, let T_{x_0} be the Taylor polynomial of degree $k-1$ at x_0+ (or at x_0-); i.e.,

$$T_{x_0}(x) := \sum_{i=0}^{k-1} \frac{1}{i!} f^{(i)}(x_0+)(x - x_0)^i.$$

Given ϵ, for δ to be prescribed, let

$$f_\epsilon := \begin{cases} T_\delta & \text{on} \quad [0,\delta], \\ f & \text{on} \quad [\delta, 1-\delta], \\ T_{1-\delta} & \text{on} \quad [1-\delta, 1]. \end{cases}$$

Then obviously $f_\epsilon \in \mathcal{M}_*^k(0,1)$ and

$$\|f - f_\epsilon\|_p \le c_p \|f - T_\delta\|_{\mathbb{L}_p[0,\delta]} + c_p \|f - T_{1-\delta}\|_{\mathbb{L}_p[1-\delta,1]}. \tag{4.3}$$

From [6, Theorem 1], it follows that, for $I = (a,b)$, $f \in \mathcal{M}_p^k(I)$, and $x_* := \frac{a+b}{2}$, we have

$$\|f - T_{x_*}\|_{\mathbb{L}_p(I)} \le c_{k,p} \omega_k(f)_{\mathbb{L}_p(I)},$$

where $\omega_k(f)_{\mathbb{L}_p(I)}$ is the kth modulus of smoothness of $f \in \mathbb{L}_p(I)$ (see section 8 for the definition), which, as is well known, has the property that $\omega_k(f)_{\mathbb{L}_p(J)} \to 0$ if $|J| \to 0$, $J \subset I$. Applying this result to the interval $(0, 2\delta) \subset (0,1)$, we obtain

$$\|f - T_\delta\|_{\mathbb{L}_p(0,\delta)} \le \|f - T_\delta\|_{\mathbb{L}_p(0,2\delta)} \le c_{k,p} \omega_k(f)_{\mathbb{L}_p(0,2\delta)} \to 0 \quad \text{as} \quad \delta \to 0.$$

Similarly,

$$\|f - T_{1-\delta}\|_{\mathbb{L}_p(1-\delta,1)} \le c_{k,p} \omega_k(f)_{\mathbb{L}_p(1-2\delta,1)} \to 0 \quad \text{as} \quad \delta \to 0. \qquad \square$$

Proof of Theorem 1.1. By Lemma 4.4, we can assume that $f \in \mathcal{M}_*^k(0,1)$. Let $\sigma \in \mathcal{S}_{N,r}$ be a spline of best $\mathbb{L}_p$-approximation to f on $(0,1)$. We need to prove that there exists a spline s such that

$$s \in \mathcal{S}_{c_0 N, r} \cap \mathcal{M}^k(0,1) \quad \text{and} \quad \|f - s\|_p \le c_1 \|f - \sigma\|_p.$$

Denote by $\{J_m\}$ the set of largest subintervals of $[0,1]$ on which σ is a polynomial of order r, and denote by $\{I_\ell\}$ the set of largest subintervals of J_m's on which $\sigma^{(k)}$ has a constant sign. Since $\sigma \in \mathcal{S}_{N,r}[0,1]$, there are at most N intervals J_m, and, on each J_m, the spline $\sigma^{(k)}$ is a polynomial of degree $r-1-k$; hence there are at most $\max(1, r-k)$ subintervals I_ℓ in each interval J_m. Thus $\{I_\ell\}$ is a partition of $[0,1]$ such that

$$[0,1] = \cup I_\ell, \quad \#\{I_\ell\} \le N \max(1, r-k),$$

and, on each I_ℓ,

$$\text{either} \quad \sigma \in \Pi_r \cap \mathcal{M}^k \qquad \text{or} \quad (-\sigma) \in (\Pi_r \setminus \Pi_k) \cap \mathcal{M}^k.$$

By Proposition 4.3, on each interval I_ℓ, there exists a spline s_ℓ such that

$$s_\ell \in \mathcal{S}_{C(k),r} \cap \mathcal{M}^k[f](I_\ell)$$

and

$$\|f - s_\ell\|_{\mathbb{L}_p(I_\ell)} \le c_1 \|f - \sigma\|_{\mathbb{L}_p(I_\ell)}. \tag{4.4}$$

Now define the spline s so that

$$s := s_\ell \quad \text{on} \quad I_\ell.$$

Relations $s_\ell \in \mathcal{S}_{C(k),r}(I_\ell)$ and $\#\{I_\ell\} \le N \max(1, r-k)$ imply that

$$s \in \mathcal{S}_{c_0 N,r}(0,1), \quad c_0 = C(k)\max(1, r-k),$$

while inclusions $s_\ell \in \mathcal{M}^k[f](I_\ell)$ with $\cup I_\ell = [0,1]$ yield

$$s \in \mathcal{M}^k[f](0,1) \subset \mathcal{M}^k(0,1).$$

Thus

$$s \in \mathcal{S}_{c_0 N,r} \cap \mathcal{M}^k(0,1).$$

Finally, to estimate the degree of approximation of f by s for $0 < p < \infty$ (modifications for $p = \infty$ are obvious), from (4.4) we obtain

$$\|f - s\|^p_{\mathbb{L}_p(0,1)} \le \sum_\ell \|f - s_\ell\|^p_{\mathbb{L}_p(I_\ell)} \le c_1^p \sum_\ell \|f - \sigma\|^p_{\mathbb{L}_p(I_\ell)} = c_1^p \|f - \sigma\|^p_{\mathbb{L}_p(0,1)},$$

i.e.,

$$E^{(k)}_{c_0 N,r}(f)_p \le c_1 E_{N,r}(f)_p. \qquad \square$$

5. k-monotone interpolation. If $\mathfrak{p} - \mathfrak{f}_*$ has many intersections (see Proposition 4.2), then the spline $\mathfrak{f}_* \in \mathcal{P}_{\mathcal{M}^k[f]}(\mathfrak{p})$ has many knots. In this case, we will modify $\mathfrak{f}_*$ into a spline s with a smaller number of knots by blending f_* with $\mathfrak{p}$. This procedure is related to the following general problem.

Problem 5.1. Given two k-monotone functions f, g on J and an interval $(a,b) \subset J$, determine whether or not there exists a k-monotone function h in $\mathcal{M}^k[f,g](a,b)$. Note that the existence of such h implies that there is a function $\widetilde{h}$ such that

$$\widetilde{h} \in \mathcal{M}^k(J) \quad \text{and} \quad \widetilde{h}(x) = \begin{cases} f(x), & x \le a, \\ g(x), & x \ge b. \end{cases}$$

We will refer to this problem as a *blending* of $f, g \in \mathcal{M}^k(J)$ on $[a,b]$. Actually, all we need is a k-monotone interpolation of data $f^{(i)}(a+), g^{(i)}(b-)$, $i = 0, \dots, k-1$, so that we consider this topic more generally.

Let

$$\boldsymbol{x} := (x_i)_{i=1}^{n+k} := \{a = x_1 \le \cdots \le x_{n+k} = b\}$$

be a sequence of interpolation knots such that $x_i < x_{i+k}$, and let

$$\boldsymbol{y} := \boldsymbol{y}(\boldsymbol{x}) := (y_i)_{i=1}^{n+k}.$$

We use the usual convention that, if some of the knots in $\boldsymbol{x}$ are repeated, then interpolation of corresponding derivatives takes place. For each $j = 1, \dots, n+k$, denote by l_j the number of points x_i such that $x_i = x_j$ with $i \le j$; i.e.,

$$l_j := l_j(\boldsymbol{x}) := \#\{i \mid \ 1 \le i \le j,\ x_i = x_j\}.$$

Note that, because of the restriction $x_i \ne x_{i+k}$, the inequality $l_j \le k$ is valid for all j.

DEFINITION 5.2. *A data sequence $(\boldsymbol{x}, \boldsymbol{y}) := (x_i, y_i)_{i=1}^{n+k}$ is called k-monotone if there exists a k-monotone function $f \in \mathcal{M}_*^k(a,b)$ such that*

$$f^{(l_j-1)}(x_j) = y_j, \quad j = 1, \dots, n+k. \tag{5.1}$$

Note that if all the knots in $\boldsymbol{x}$ are distinct, then the sequence $(\boldsymbol{x}, \boldsymbol{y})$ is k-monotone if $f(x_i) = y_i$, $j = 1, \dots, n+k$, for some $f \in \mathcal{M}_*^k(a,b)$. Also, if $l_j = k$ for some j, then $f^{(l_j-1)}(x_j) = f^{(k-1)}(x_j)$ is understood as $f^{(k-1)}(x_j+)$ or $f^{(k-1)}(x_j-)$.

Since

$$f \in \mathcal{M}^k \quad \Leftrightarrow \quad [t_i, \dots, t_{i+k}]f \ge 0 \quad \forall\ (t_i),$$

where not all t_i's are the same, one must necessarily have for a k-monotone sequence $(\boldsymbol{x}, \boldsymbol{y})$

$$[x_i, \dots, x_{i+k}]\boldsymbol{y} \ge 0.$$

If $k = 1$ or 2 (i.e., in the case of monotone or convex interpolation), this condition is sufficient as well. However, it is *not* sufficient if $k \ge 3$, as the following example shows.

Example 5.3. The data set

$\boldsymbol{x}$	$\boldsymbol{y}$	δ^1	δ^2	δ^3
-5	-77			
		25		
-3	-27		-3	
		13		0
-1	-1		-3	
		1		1
1	1		3	
		13		0
3	27		3	
		25		
5	77			

has nonnegative divided differences of order 3, but, at the same time,

$$\begin{aligned}
&[-5,-3,-1,0]\boldsymbol{y} + [0,1,3,5] \\
&\quad = \ -\frac{1}{5}[-5,-3,-1]\boldsymbol{y} - \frac{1}{15}[-3,-1]\boldsymbol{y} - \frac{1}{15}[-1]\boldsymbol{y} + \frac{1}{15}[0]\boldsymbol{y} \\
&\qquad\quad + \frac{1}{5}[1,3,5]\boldsymbol{y} \quad - \frac{1}{15}[1,3]\boldsymbol{y} \quad + \frac{1}{15}[1]\boldsymbol{y} - \frac{1}{15}[0]\boldsymbol{y} \\
&\quad = \ \frac{1}{5}\cdot 6 - \frac{1}{15}\cdot 26 + \frac{1}{15}\cdot 2 = -\frac{6}{15} < 0.
\end{aligned}$$

Hence there is no 3-monotone function passing through $(\boldsymbol{x}, \boldsymbol{y})$.

Denote by

$$\boldsymbol{v} := \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}) := (v_i)_1^n, \quad v_i := [x_i, \dots, x_{i+k}]\boldsymbol{y},$$

the sequence of divided differences of $\boldsymbol{y}(\boldsymbol{x})$, and denote by

$$\mathfrak{M} := \mathfrak{M}(\boldsymbol{x}) := \left(\frac{1}{k!}M_i\right), \quad M_i(t) := k[x_i, \dots, x_{i+k}](\cdot - t)_+^{k-1},$$

the sequence of the B-splines of order k with the knot sequence $\boldsymbol{x}$. Recall that $\operatorname{supp} M_i = [x_i, x_{i+k}]$, $M_i \ge 0$, $\int M_i = 1$, and, for any $f \in \mathbb{C}^k(a,b)$ (in fact, the condition $f \in \mathbb{W}_1^k(a,b)$ is sufficient),

$$[x_i, \dots, x_{i+k}]f = \frac{1}{k!}\int_a^b M_i(t) f^{(k)}(t)\, dt.$$

Notice that if a k-monotone function f belongs to $\mathbb{C}^k$, then $f^{(k)} \ge 0$. Thus, to check whether the data sequence (x_i, y_i) is k-monotone, one needs to form the sequence of divided differences (v_i) and check whether there is a nonnegative function λ such that

$$v_i = \frac{1}{k!}\int M_i(t)\lambda(t)\, dt.$$

The last problem is the so-called *Markov moment problem*, which we discuss in the next section.

6. Markov moment problem and k-monotone interpolation. Let $\mathcal{U} := (u_i)_{i=1}^n$ be a sequence of continuous linearly independent real-valued functions on $I = (a,b)$, and let $\boldsymbol{v} := (v_i)_{i=1}^n$ be a sequence of real numbers.

DEFINITION 6.1. *A sequence $\boldsymbol{v} \in \mathbb{R}^n$ is called a moment sequence with respect to $\mathcal{U}$ if, for some bounded nondecreasing function μ, it admits the representation*

$$v_i = \int_a^b u_i(t)\, d\mu(t), \quad 1 \le i \le n.$$

LEMMA 6.2. *A data sequence $(\boldsymbol{x}, \boldsymbol{y})$ is k-monotone if and only if the sequence of divided differences $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y})$ is a moment sequence with respect to $\mathfrak{M}(\boldsymbol{x})$, the sequence of B-splines.*

Proof. By Lemma 3.1(3), $f \in \mathcal{M}_*^k(a,b)$ can be represented as

$$f(x) = p(x) + \frac{1}{k!}\int_a^b k(x-t)_+^{k-1}\, d\mu(t), \tag{6.1}$$

where $p \in \Pi_k$ and μ is a bounded nondecreasing function. If $f|_{\boldsymbol{x}} = \boldsymbol{y}$, then

$$v_i := [x_i, \dots, x_{i+k}]\boldsymbol{y} = [x_i, \dots, x_{i+k}]f = \frac{1}{k!}\int_a^b M_i(t)\, d\mu(t),$$

i.e., $\boldsymbol{v}$ is a moment sequence with respect to $\mathfrak{M}$.

Conversely, if for the sequences $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y})$ and $\mathfrak{M}(\boldsymbol{x})$ there exists a bounded nondecreasing function μ such that

$$v_i = \frac{1}{k!}\int_a^b M_i(t)\, d\mu(t), \quad i = 1, \dots, n,$$

then, for any $p \in \Pi_k$, the function f defined by (6.1) is in $\mathcal{M}_*^k$ and satisfies

$$[x_i, \dots, x_{i+k}]f = v_i := [x_i, \dots, x_{i+k}]\boldsymbol{y}, \quad i = 1, \dots, n. \tag{6.2}$$

Finally, in (6.1), we can choose $p \in \Pi_k$ so that the equality in (5.1) holds for $j = 1, \dots, k$, and that, together with (6.2), implies successively that it is also true for $j = k+1, \dots, n+k$; hence the sequence (x_i, y_i) is k-monotone. ∎

Now we need a result from the theory of moments which gives a characterization of the moment sequences.

DEFINITION 6.3. *A sequence $\boldsymbol{v} \in \mathbb{R}^n$ of real numbers is called positive with respect to $\mathcal{U} = (u_i)_{i=1}^n$ (recall that $\mathcal{U}$ is a sequence of continuous linearly independent real-valued functions on $[a, b]$) if*

$$\sum_{i=1}^{n} a_i u_i(t) \geq 0, \quad a \leq t \leq b, \qquad \Rightarrow \qquad \sum_{i=1}^{n} a_i v_i \geq 0.$$

THEOREM 6.4 (Krein and Nudelman [7, Theorem 3.1.1, p. 58], [8]). *Let $\mathcal{U} := (u_i)_{i=1}^n$ be a sequence of continuous linearly independent real-valued functions on $I = [a, b]$ with the property that there exists a strictly positive polynomial $p \in \operatorname{span} \mathcal{U}$. A sequence $\boldsymbol{v} \in \mathbb{R}^n$ is a moment sequence with respect to $\mathcal{U}$ if and only if $\boldsymbol{v}$ is positive with respect to $\mathcal{U}$.*

Since $\operatorname{span} \mathfrak{M}(\boldsymbol{x})$ contains constants, we may combine this theorem with Lemma 6.2 to obtain the following criterion for k-monotonicity of data.

COROLLARY 6.5. *A data sequence $(\boldsymbol{x}, \boldsymbol{y})$ is k-monotone if and only if the sequence of divided differences $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y})$ is positive with respect to $\mathfrak{M}(\boldsymbol{x})$, i.e., if and only if*

$$\sum_{i=1}^{n} a_i M_i(t) \geq 0 \quad \Rightarrow \quad \sum_{i=1}^{n} a_i v_i \geq 0, \quad v_i = [x_i, \dots, x_{i+k}]\boldsymbol{y}.$$

7. Blending of k-monotone functions. In this section, we will give a partial solution to Problem 5.1. Namely, in Proposition 7.3, we prove that, provided f and g have sufficiently many points of intersection, a function $h \in \mathcal{M}^k[f, g]$ exists.

We need two auxiliary statements.

The following lemma is a particular case of Lemma 3.2 in Beatson [1] concerning the spline blending. Actually, we will use a more detailed statement which is formulated within the proof of Proposition 7.3.

LEMMA 7.1 (Beatson [1]). *Let $k \in \mathbb{N}$, $n = 2k^2$, and let $p \in \Pi_k$ be a nonnegative polynomial on $[a, b]$. Then, for any knot sequence*

$$\mathbf{t} := \{a = t_0 \leq t_1 \leq \cdots \leq t_n < t_{n+1} = b\},$$

there exists a nonnegative spline $s_2 \in \mathcal{S}_{\mathbf{t},k}(\mathbb{R})$ (i.e., s_2 is a spline of order r on the knot sequence $\mathbf{t}$) such that

$$s_2 \equiv 0 \quad \text{on} \quad (-\infty, a], \qquad 0 \leq s_2 \leq p \quad \text{on} \quad [a, b], \qquad s_2 = p \quad \text{on} \quad [b, \infty).$$

The next statement is a well-known property of divided differences.

LEMMA 7.2. *Let $(x_j)_{j=1}^{n+k}$ be any nondecreasing sequence such that $x_j < x_{j+k}$. Then, for any subsequence $(x_{i_0}, \dots, x_{i_k})$ of length $k+1$, there exist coefficients ν_j such that, for any continuous f (which is differentiable at the repeated knots),*

$$[x_{i_0}, \dots, x_{i_k}]f = \sum_{j=1}^{n} \nu_j [x_j, \dots, x_{j+k}]f.$$

PROPOSITION 7.3. *For $k \in \mathbb{N}$ and $n = 2k^2$, let $f, g \in \mathcal{M}_*^k(a,b)$ be such that*

$$f(t_j) = g(t_j) \quad on \quad \{a = t_0 < t_1 < \cdots < t_n < t_{n+1} = b\}.$$

Then there exists a function $h \in \mathcal{M}_*^k(a,b)$ such that

$$h^{(l)}(a+) = f^{(l)}(a+), \quad h^{(l)}(b-) = g^{(l)}(b-), \quad l = 0, \dots, k-1.$$

Note that the condition that all points t_i in the statement of Proposition 7.3 are distinct is not essential and is used here only in order to simplify the exposition.

Proof. Let us introduce two sequences $\boldsymbol{x} = (x_i)_{i=1}^{n+2k}$ and $\boldsymbol{y} = (y_i)_{i=1}^{n+2k}$:

$$x_j := \begin{cases} a, & 1 \le j \le k, \\ t_{j-k}, & k+1 \le j \le n+k, \\ b, & n+k+1 \le j \le n+2k, \end{cases}$$

and

$$y_j := \begin{cases} f^{(j-1)}(a+), & 1 \le j \le k, \\ f(x_j) = g(x_j), & k+1 \le j \le n+k, \\ g^{(j-n-k-1)}(b-), & n+k+1 \le j \le n+2k. \end{cases}$$

It is convenient to arrange this data set $(\boldsymbol{x}, \boldsymbol{y})$ as follows:

$$\begin{array}{cccccccccccc}
\boldsymbol{y}_1 & & & & & & & & & & & \\
\| & & & & & & & & & & & \\
(y_j)_{j=1}^{n+k+1} \to & f(a) & \dots & f^{(k-1)}(a) & f(x_{k+1}) & \dots & f(x_{n+k}) & & f(b) & & & \\
& \uparrow & & \uparrow & \uparrow & & \uparrow & & \uparrow & & & \\
& x_1 = \cdots = x_k & < & x_{k+1} & \le \cdots \le & x_{n+k} & < & b = & x_{n+k+1} = \cdots = & x_{n+2k} & & \\
& \downarrow & & \downarrow & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\
& g(a) & & g(x_{k+1}) & \dots & g(x_{n+k}) & & g(b) & \dots & g^{(k-1)}(b) & \leftarrow (y_j)_{j=k}^{n+2k} & \\
& & & & & & & & & & \| & \\
& & & & & & & & & & \boldsymbol{y}_2 &
\end{array}$$

Set

$$\begin{aligned} \boldsymbol{x}_* &:= (x_1, \dots, x_k, x_{n+k+1}, \dots, x_{n+2k}) := (\overbrace{a, \dots, a}^{k}, \overbrace{b, \dots, b}^{k}), \\ \boldsymbol{y}_* &:= (y_1, \dots, y_k, y_{n+k+1}, \dots, y_{n+2k}) \\ &:= \left(f(a), \dots, f^{(k-1)}(a), g(b), \dots, g^{(k-1)}(b)\right). \end{aligned} \tag{7.1}$$

We need to interpolate $\boldsymbol{y}_*$ on $\boldsymbol{x}_*$ by a k-monotone function h. Denote by

$$\mathfrak{M}(\boldsymbol{x}_*) =: (B_i)_{i=1}^k, \quad \boldsymbol{v}(\boldsymbol{x}_*, \boldsymbol{y}_*) =: (w_i)_{i=1}^k$$

the sequences of the B-splines and of divided differences, respectively, which correspond to $(\boldsymbol{x}_*, \boldsymbol{y}_*)$. By Corollary 6.5, existence of a k-monotone interpolant h to the data (7.1) will follow if we show that

$$\sum_{i=1}^k a_i B_i(t) \ge 0 \quad \Rightarrow \quad \sum_{i=1}^k a_i w_i \ge 0. \tag{7.2}$$

We start with some preliminaries.

(1) Let $(v_j)_{j=1}^{n+k}$ and $(M_j)_{j=1}^{n+k}$ be the sequences of divided differences and B-splines, respectively, constructed with respect to the entire set $(x_j, y_j)_{j=1}^{n+2k}$. Consider two sets of the following subsequences:

$$\begin{array}{llll} \boldsymbol{x}_1 := (x_j)_{j=1}^{n+k+1}, & \boldsymbol{y}_1 := (y_j)_{j=1}^{n+k+1}, & \boldsymbol{v}_1 := (v_j)_{j=1}^{n+1}, & \mathfrak{M}_1 := (M_j)_{j=1}^{n+1}; \\ \boldsymbol{x}_2 := (x_j)_{j=k}^{n+2k}, & \boldsymbol{y}_2 := (y_j)_{j=k}^{n+2k}, & \boldsymbol{v}_2 := (v_j)_{j=k}^{n+k}, & \mathfrak{M}_2 := (M_j)_{j=k}^{n+k}. \end{array} \tag{7.3}$$

By assumption, k-monotone f interpolates $\boldsymbol{y}_1$ on $\boldsymbol{x}_1$, and k-monotone g interpolates $\boldsymbol{y}_2$ on $\boldsymbol{x}_2$; thus both sets of data $(\boldsymbol{x}_\nu, \boldsymbol{y}_\nu)$, $\nu = 1, 2$, are k-monotone. Then Corollary 6.5 implies that

$$\boldsymbol{v}_\nu \text{ is positive with respect to } \mathfrak{M}_\nu, \quad \nu = 1, 2. \tag{7.4}$$

(2) Since (v_j), (M_j) are divided differences of certain functions on $\boldsymbol{x}$, while (w_i), (B_i) are divided differences of the same functions on $\boldsymbol{x}_* \subset \boldsymbol{x}$, by Lemma 7.2 there exist expansions

$$w_i = \sum_{j=1}^{n+k} c_{ij} v_j, \quad B_i(x) = \sum_{j=1}^{n+k} c_{ij} M_j(x)$$

with the same coefficients (c_{ij}) in both of these equations. This implies that, for any $(a_i)_{i=1}^k \subset \mathbb{R}$, the expansions

$$\sum_{i=1}^{k} a_i B_i(x) = \sum_{j=1}^{n+k} c_j M_j(x), \quad \sum_{i=1}^{k} a_i w_i = \sum_{j=1}^{n+k} c_j v_j$$

have the same coefficients $c_j = \sum_{i=1}^k a_i c_{ij}$.

(3) The B-splines $(B_i) \in \mathfrak{M}(\boldsymbol{x}_*)$ have the form

$$\begin{aligned} B_i(t) &:= k[\overbrace{a, \dots, a}^{k+1-i}, \overbrace{b, \dots, b}^{i}](\cdot - t)_+^{k-1} \\ &= \frac{k}{(b-a)^k} \binom{k-1}{i-1} (t-a)^{i-1} (b-t)^{k-i}, \quad i = 1, \dots, k, \end{aligned}$$

i.e., they are Bernstein basis polynomials of order k, so that

$$\sum a_i B_i \in \Pi_k \quad \forall\ (a_i).$$

Now let us prove (7.2). Suppose that, for some sequence (a_i),

$$p_a(x) := \sum_{i=1}^{k} a_i B_i(x) \geq 0.$$

Since p_a is a polynomial of order k,

$$p_a(x) := \sum_{i=1}^{k} a_i B_i(x) = \sum_{j=1}^{n+k} c_j M_j(x) \geq 0, \quad \text{and} \quad n \geq 2k^2,$$

the method of the proof of Beatson's smoothing lemma [1, Lemma 3.2] shows that there is an index l, $k \le l \le n+1$, such that

$$s_1(x) := \sum_{j=1}^{l} c_j M_j(x) \ge 0, \quad s_2(x) := \sum_{j=l+1}^{n+k} c_j M_j(x) \ge 0. \tag{7.5}$$

(We will not repeat Beatson's argument here and mention only that the sign variation diminishing property of B-spline series (see [4, section 5.10], for example) as well as their finite support are used.) From the definitions in (7.3), it follows that $s_\nu \in \mathfrak{M}_\nu$, $\nu = 1, 2$, which allows us to conclude that, since $\boldsymbol{v}_\nu$ are positive with respect to $\mathfrak{M}_\nu$ (see (7.4)), (7.5) implies

$$\sum_{j=1}^{l} c_j v_j \ge 0, \quad \sum_{j=l+1}^{n+k} c_j v_j \ge 0.$$

Finally,

$$\sum_{i=1}^{k} a_i w_i = \sum_{j=1}^{n+k} c_j v_j = \sum_{i=1}^{l} c_j v_j + \sum_{j=l+1}^{n+k} c_j v_j \ge 0.$$

Hence (7.2) is proved, and the proof of the proposition is now complete. □

Now, having proved the existence of a function $h \in \mathcal{M}^k[f, g]$, we may use Lemma 3.4 to derive the existence of a spline $z \in \mathcal{M}^k[f, g]$.

COROLLARY 7.4. *For $k \in \mathbb{N}$, $n = 2k^2$, let $f, g \in \mathcal{M}^k_*(a, b)$ be such that*

$$f(t_j) = g(t_j) \quad \text{on} \quad \{a = t_0 < t_1 < \cdots < t_n < t_{n+1} = b\}.$$

Then there exists a spline z such that

$$z \in \mathcal{S}_{k',k} \cap \mathcal{M}^k[f, g].$$

Note that, for $k = 1$ or 2, that is, for monotone or convex functions f and g, a procedure of k-monotone blending of f and g is quite evident geometrically.

8. Auxiliary Whitney-type estimates. In this section, we give some Whitney-type estimates for approximation of polynomials $\mathfrak{p} \in \Pi_r$ by splines and polynomials of degree k.

As usual, $\omega_k(f, \delta, I)_p$ denotes the kth modulus of smoothness of f with the step δ on the interval I,

$$\omega_k(f, \delta, I)_p := \sup_{0<h\le\delta} \left\| \Delta_h^k(f, \cdot, I) \right\|_{\mathbb{L}_p(I)},$$

where $\Delta_h^k(f, x, I)$ is the kth forward difference,

$$\Delta_h^k(f, x, I) := \begin{cases} \sum_{i=0}^{k} \binom{k}{i} (-1)^{k-i} f(x + ih) & \text{if } [x, x + kh] \subset I, \\ 0 & \text{otherwise.} \end{cases}$$

It is also convenient to denote

$$\omega_k(f)_p := \omega_k(f)_{\mathbb{L}_p(I)} := \omega_k(f, |I|, I)_p.$$

LEMMA 8.1. *Let $k, r \in \mathbb{N}$, $0 < p \le \infty$, $I = (a, b)$, and $\mathfrak{p} \in \Pi_r$, and let s be a spline of order k with at most $C(k)$ pieces in I (i.e., $s \in \mathcal{S}_{C(k),k}$). Then*

$$\|\mathfrak{p} - s\|_p \ge c_{p,r,k}\omega_k(\mathfrak{p})_p.$$

Proof. Let J be a largest subinterval of I between two successive knots of s (and hence $|J|/|I| \ge 1/C(k)$), and let $\mathfrak{q} \in \Pi_k$ be the restriction of s to J. Then, using Whitney's inequality,

$$E(\mathfrak{p}, \Pi_k)_{\mathbb{L}_p(I)} \overset{p,k}{\sim} \omega_k(\mathfrak{p})_{\mathbb{L}_p(I)}, \tag{8.1}$$

and the Markov-type inequality (see [4, (4.2.10) and (4.2.16)])

$$\|\mathfrak{p}\|_{\mathbb{L}_p(I)} \le c_{p,r}(|I|/|J|)^{r-1+1/p}\|\mathfrak{p}\|_{\mathbb{L}_p(J)},$$

we find

$$\begin{aligned}\|\mathfrak{p} - s\|_{\mathbb{L}_p(I)} &\ge \|\mathfrak{p} - \mathfrak{q}\|_{\mathbb{L}_p(J)} \ge c_{p,r}(|J|/|I|)^{r-1+1/p}\|\mathfrak{p} - \mathfrak{q}\|_{\mathbb{L}_p(I)} \\ &\ge c_{p,r,k}E(\mathfrak{p}, \Pi_k)_{\mathbb{L}_p(I)} \ge c'_{p,r,k}\omega_k(\mathfrak{p})_p. \qquad \square\end{aligned}$$

LEMMA 8.2. *Let $k, r \in \mathbb{N}$, $0 < p \le \infty$, $I = (a, b)$, and $\mathfrak{p} \in \Pi_r$, and let $l_k(\mathfrak{p})$ be the Lagrange polynomial of order k interpolating $\mathfrak{p}$ at any k (not necessarily distinct) points inside I. Then*

$$\|\mathfrak{p} - l_k(\mathfrak{p})\|_p \le c_{p,r,k}\omega_k(\mathfrak{p})_p.$$

Proof. Taking into account Lebesgue's inequality

$$\|\mathfrak{p} - l_k(\mathfrak{p})\|_p \le \left(\sup_{\mathfrak{q}\in\Pi_r} \frac{\|l_k(\mathfrak{q})\|_p}{\|\mathfrak{q}\|_p} + 1\right) E(\mathfrak{p}, \Pi_k)_p$$

and Whitney's inequality (8.1), it suffices to prove that

$$\|l_k(\mathfrak{q})\|_p \le c_{p,r,k}\|\mathfrak{q}\|_p \quad \forall\ \mathfrak{q} \in \Pi_r.$$

We make use of Markov's inequality

$$\|\mathfrak{q}^{(k)}\|_\infty \le c_{p,r,k}|I|^{-k-1/p}\|\mathfrak{q}\|_p$$

and the well-known error bound for the Lagrange interpolation

$$\|f - l_k(f)\|_\infty \le c_k|I|^k\|f^{(k)}\|_\infty$$

to obtain

$$\|\mathfrak{q} - l_k(\mathfrak{q})\|_p \le |I|^{1/p}\|\mathfrak{q} - l_k(\mathfrak{q})\|_\infty \le c_k|I|^{1/p}|I|^k\|\mathfrak{q}^{(k)}\|_\infty \le c_{p,r,k}\|\mathfrak{q}\|_p. \qquad \square$$

LEMMA 8.3. *Let $k \in \mathbb{N}$ and $f \in \mathcal{M}^k(a, b)$, and let $l_k(f, x; x_1, \dots, x_k)$ be the Lagrange (Hermite–Taylor) polynomial of degree $\le k-1$ interpolating f at the points x_i, $1 \le i \le k$, where $a =: x_0 < x_1 \le \cdots \le x_k < x_{k+1} := b$. Then*

$$(-1)^{k-i}(f(x) - l_k(f, x; x_1, \dots, x_k)) \ge 0, \quad x \in (x_i, x_{i+1}), \quad i = 0, \dots, k.$$

In other words, $f - l_k$ changes sign at $x_1, \dots, x_k$.

Proof. First, if all the points x_i, $1 \le i \le k$, are distinct, this is Theorem 5 in Bullen [3].

If some of x_i (but not all) coincide, the statement of the lemma is a consequence of the following result which follows from [4, Theorem 4.6.3]: *For a given $f \in \mathbb{C}^{(k-2)}(a,b)$, the Lagrange–Hermite polynomial $l_k(X) = l_k(f,x;x_1,\dots,x_k)$ is a continuous function of $X = (x_1,\dots,x_k)$ at each point $X^* = (x_1^*,\dots,x_k^*) \in (a,b)^k$ such that not all x_i^*, $i = 1,\dots,k$, are the same.*

If all points coincide, i.e., $x_1 = \cdots = x_k = \xi$, the lemma follows from the following statement, which can be proved by induction on k: *Let $k \in \mathbb{N}$, $f \in \mathcal{M}^k(a,b)$, and $\xi \in (a,b)$. If t_k is a Taylor polynomial of degree $\le k-1$ for f at ξ, i.e., $t_k^{(i)}(\xi) = f^{(i)}(\xi\pm)$ for $i = 0,\dots,k-1$ (or, more precisely, $t_k^{(i)}(\xi) = f^{(i)}(\xi)$ for $i = 0,\dots,k-2$ and $t_k^{(k-1)}(\xi)$ is either $f^{(k-1)}(\xi+)$ or $f^{(k-1)}(\xi-)$), then*

$$f(x) - t_k(x) \ge 0,\ x \in (\xi,b), \quad \text{and} \quad (-1)^k(f(x) - p_k(x)) \ge 0,\ x \in (a,\xi). \qquad \square$$

The following is an immediate corollary of Lemma 8.3.

COROLLARY 8.4. *For $k \in \mathbb{N}$, $f \in \mathcal{M}_*^k(a,b)$, and a set of interpolation points $\{a = x_0 \le \cdots \le x_k = b\}$, let*

$$l_k := l_k(f; x_0 \dots, x_{k-1}) \quad \textit{and} \quad \widetilde{l}_k := \widetilde{l}_k(f, x_1 \dots, x_k)$$

be two Lagrange (Hermite–Taylor) interpolants to f on the given sets. Then f lies between l_k and $\widetilde{l}_k$ on $[a,b]$; i.e.,

$$\min\{l_k, \widetilde{l}_k\} \le f \le \max\{l_k, \widetilde{l}_k\}.$$

LEMMA 8.5. *Let $k, r \in \mathbb{N}$, $0 < p \le \infty$, $I = (a,b)$, $\mathfrak{p} \in \Pi_r \cap \mathcal{M}^k$, $0 \le \mu \le k-1$, and let $g \in \mathcal{M}^k$ be a function such that*

$$g^{(i)}(a) = \mathfrak{p}^{(i)}(a), \quad i = 0,\dots,\mu, \tag{8.2}$$

and

$$g^{(i)}(b) = \mathfrak{p}^{(i)}(b), \quad i = 0,\dots,k-\mu-1. \tag{8.3}$$

(Here, in the cases $\mu = 0$ and $\mu = k-1$, $g^{(k-1)}(b)$ and $g^{(k-1)}(a)$ are understood as $g^{(k-1)}(b-)$ and $g^{(k-1)}(a+)$, respectively.) Then

$$\|\mathfrak{p} - g\|_p \le c_{p,r,k}\omega_k(\mathfrak{p})_p.$$

Proof. Consider the following Lagrange (Hermite–Taylor) polynomials of order k on $[a,b]$:

$$l_k := l_k(\mathfrak{p}; \overbrace{a,\dots,a}^{\mu+1}, \overbrace{b,\dots,b}^{k-\mu-1}) \quad \text{and} \quad \widetilde{l}_k := \widetilde{l}_k(\mathfrak{p}; \overbrace{a,\dots,a}^{\mu}, \overbrace{b,\dots,b}^{k-\mu}).$$

By Corollary 8.4, both k-monotone functions $\mathfrak{p}$ and g lie between l_k and $\widetilde{l}_k$ in $[a,b]$, i.e.,

$$\min\{l_k, \widetilde{l}_k\} \le \min\{\mathfrak{p}, g\} \le \max\{\mathfrak{p}, g\} \le \max\{l_k, \widetilde{l}_k\}.$$

Therefore,

$$\|g - \mathfrak{p}\|_p \le \|l_k - \widetilde{l}_k\|_p \le c_p\|l_k - \mathfrak{p}\|_p + c_p\|\mathfrak{p} - \widetilde{l}_k\|_p \le c_{p,r,k}\omega_k(\mathfrak{p})_p,$$

where the last inequality follows from Lemma 8.2. ∎

In our proof, we need a slightly stronger statement in the case for $\mu = 0$.

LEMMA 8.6. *Let $k, r \in \mathbb{N}$, $0 < p \le \infty$, $I = (a, b)$, and $\mathfrak{p} \in \Pi_r \cap \mathcal{M}^k$, and let $h \in \mathcal{M}^k$ be a function such that*

$$h(a) = \mathfrak{p}(a) \quad \text{and} \quad h^{(i)}(b) = \mathfrak{p}^{(i)}(b),\ i = 0, \dots, k-2, \quad h^{(k-1)}(b-) \le \mathfrak{p}^{(k-1)}(b). \tag{8.4}$$

Then

$$\|\mathfrak{p} - h\|_p \le c_{p,r,k}\omega_k(\mathfrak{p})_p. \tag{8.5}$$

Proof. First, assume that there exists $\delta > 0$ such that $\mathfrak{p} \in \mathcal{M}^k(a, b+\delta)$, and set

$$g = \begin{cases} h & \text{on} \quad [a, b), \\ \mathfrak{p} & \text{on} \quad [b, b+\delta]. \end{cases}$$

Then g is k-monotone on $[a, b+\delta]$ and satisfies all other assumptions of Lemma 8.5 (with $\mu = 0$); hence

$$\|g - \mathfrak{p}\|_{\mathbb{L}_p[a,b+\delta]} \le c_{p,r,k}\omega_k(\mathfrak{p})_{\mathbb{L}_p[a,b+\delta]}.$$

Letting $\delta \to 0$, we obtain

$$\|h - \mathfrak{p}\|_{\mathbb{L}_p[a,b]} \le \lim_{\delta\to 0} \|g - \mathfrak{p}\|_{\mathbb{L}_p[a,b+\delta]} \le c_{p,r,k} \lim_{\delta\to 0} \omega_k(\mathfrak{p})_{\mathbb{L}_p[a,b+\delta]} = c_{p,r,k}\omega_k(\mathfrak{p})_{\mathbb{L}_p[a,b]}.$$

If, for any $\delta > 0$, $\mathfrak{p} \notin \mathcal{M}^k(a, b+\delta)$, one can replace $\mathfrak{p}$ by

$$\widetilde{\mathfrak{p}}(x) := \mathfrak{p}(x) + \epsilon(x-a)(x-b)^{k-1}.$$

Then $\widetilde{\mathfrak{p}} \in \Pi_{\max\{r,k+1\}} \cap \mathcal{M}^k(a, b+\Delta)$ for some $\Delta > 0$,

$$\widetilde{\mathfrak{p}}(a) = \mathfrak{p}(a), \quad \widetilde{\mathfrak{p}}^{(i)}(b) = \mathfrak{p}^{(i)}(b),\ 0 \le i \le k-2,$$

and

$$\widetilde{\mathfrak{p}}^{(k-1)}(b) = \mathfrak{p}^{(k-1)}(b) + (k-1)!\,\epsilon(b-a) \ge \mathfrak{p}^{(k-1)}(b) \ge h^{(k-1)}(b-).$$

Now using the same argument as above and letting $\epsilon \to 0$ and $\Delta \to 0$ complete the proof of the lemma. ∎

9. Proof of Proposition 4.2. The following statement summarizes the results of sections 5–8.

PROPOSITION 9.1. *Let $k \in \mathbb{N}$, $n = 2k^2$, $0 < p \le \infty$, $I = (a, b)$, and $\mathfrak{p} \in \Pi_r \cap \mathcal{M}^k$, and let $g_* \in \mathcal{S}_{C(k),k} \cap \mathcal{M}^k_*$ be such that*

$$g_*(t_j) = \mathfrak{p}(t_j) \quad \text{on} \quad \{a = t_0 < t_1 < \cdots < t_n < t_{n+1} = b\}.$$

Then there exists a spline z such that

$$z \in \mathcal{S}_{k',k} \cap \mathcal{M}^k[g_*, \mathfrak{p}]$$

and

$$\|\mathfrak{p} - z\|_p \le c_2\|\mathfrak{p} - g_*\|_p, \quad c_2 = c_2(p, r, k). \tag{9.1}$$

Proof. First, Corollary 7.4 implies that there exists a spline $z \in \mathcal{S}_{k',k} \cap \mathcal{M}^k[g_*, \mathfrak{p}]$. Now, since z satisfies condition (8.4) of Lemma 8.6 (which follows from the definition of the class $\mathcal{M}^k[g_*, \mathfrak{p}]$ and the fact that $g_*(a) = \mathfrak{p}(a)$), we have the estimate

$$\|\mathfrak{p} - z\|_p \le c_{p,r,k}\omega_k(\mathfrak{p})_p.$$

On the other hand, for $g_* \in \mathcal{S}_{C(k),k}$, Lemma 8.1 yields

$$\|\mathfrak{p} - g_*\|_p \ge c_{p,r,k}\omega_k(\mathfrak{p})_p.$$

Combining both estimates, we obtain (9.1). □

Remark 9.2. Applying Proposition 9.1 to $\widetilde{\mathfrak{p}}(t) := (-1)^k\mathfrak{p}(-t)$ and $\widetilde{g}_*(t) := (-1)^k g_*(-t)$, we conclude that there also exists a spline $\widetilde{z} \in \mathcal{S}_{k',k} \cap \mathcal{M}^k[\mathfrak{p}, g_*]$ for which (9.1) is valid.

We also need the following elementary statement.

LEMMA 9.3. *Let $(x_j)_{j=1}^{\infty}$ be such that $x_i \ne x_j$ if $i \ne j$, and $\lim_{j\to\infty} x_j = L$, and let, for some $k \ge 2$, f be $(k-2)$ times continuously differentiable in some ϵ-neighborhood of L and have one-sided $(k-1)$st derivatives at L. If $f(x_j) = 0$ for all j, then $f^{(i)}(L) = 0$ for $i = 0, \dots, k-2$ and either $f^{(k-1)}(L+) = 0$ or $f^{(k-1)}(L-) = 0$.*

Proof of Proposition 4.2. If $0 < p < \infty$, let $\mathfrak{f}_*$ be a best $\mathbb{L}_p$-approximant to $\mathfrak{p} \in \Pi_r \cap \mathcal{M}^k$ from the set $\mathcal{M}^k[f]$ whose existence is guaranteed by Lemma 3.3, and so Lemma 3.8 is valid. If $p = \infty$, we choose $\mathfrak{f}_*$ to be a best $\mathbb{L}_\infty$-approximant to $\mathfrak{p}$ from the set $\mathcal{M}^k[f]$ which satisfies Lemma 3.9.

We need to prove that there exists a spline s such that

$$s \in \mathcal{S}_{C(k),r} \cap \mathcal{M}^k[f] \tag{9.2}$$

and

$$\|\mathfrak{p} - s\|_p \le c_2\|\mathfrak{p} - \mathfrak{f}_*\|_p. \tag{9.3}$$

Lemmas 3.8 and 3.9 imply that, on any interval (c, d) where the difference $\mathfrak{f}_*(x) - \mathfrak{p}(x)$ has exactly $m-1$ distinct zeros, we have

$$\mathfrak{f}_* \in \mathcal{S}_{mk',k}, \quad k' = \lfloor k/2 \rfloor + 1. \tag{9.4}$$

Denote by $\mathfrak{Z}$ the set of all zeros of the function $\mathfrak{f}_* - \mathfrak{p}$, i.e.,

$$\mathfrak{Z} := \{x \in [a, b] \mid \mathfrak{f}_*(x) = \mathfrak{p}(x)\},$$

and let $\mathfrak{Z}^*$ be the set of all limit points of $\mathfrak{Z}$. Also, let $\#\mathfrak{Z}$ denote the cardinality of $\mathfrak{Z}$. (Note that the set $\mathfrak{Z}$ does not take into account multiplicity of zeros. This is not essential and is done only to simplify the exposition.)

The proof is quite transparent. If $\mathfrak{Z}$ consists of only a few (less than $4k^2 + 4$) points, (9.4) implies that $\mathfrak{f}_*$ has to be in $\mathcal{S}_{C(k),k}$, and so there is nothing to prove. If $\#\mathfrak{Z}$ is not less than $4k^2 + 4$ but is finite, we use Proposition 9.1 to blend $\mathfrak{f}_*$ and $\mathfrak{p}$ on intervals containing the first and the last $2k^2 + 2$ points from $\mathfrak{Z}$ (and hence $\mathfrak{f}_*$, which has many "knots" between these intervals, is replaced there by the polynomial $\mathfrak{p}$). Finally, if $\mathfrak{Z}$ is an infinite set, the set $\mathfrak{Z}^*$ is necessarily not empty and connected. Hence $\mathfrak{Z}^*$ is a closed subinterval of (or a point in) $[a, b]$. We will show that $\mathfrak{f}_* \equiv \mathfrak{p}$ on $\mathfrak{Z}^*$, and so it will remain to apply the above-mentioned argument in the case in which $\#\mathfrak{Z} < \infty$ to the set $[a, b] \setminus \mathfrak{Z}^*$ which is a union of at most two intervals.

We now fill in the details and consider the following three cases.

Case 1. $\#\mathfrak{Z} < 4k^2 + 4$. According to (9.4),

$$\mathfrak{f}_* \in \mathcal{S}_{C(k),k}, \quad C(k) \le (4k^2+4)k',$$

so we let $s = \mathfrak{f}_*$.

Case 2. $4k^2+4 \le \#\mathfrak{Z} < \infty$. Denote by $I_\nu := [a_\nu, b_\nu]$, $\nu = 1, 2$, the smallest closed subintervals of $[a, b]$ which contain the first and the last $2k^2+2$ points of $\mathfrak{Z}$, respectively (i.e., $a_1 = \min(\mathfrak{Z})$ and $b_2 = \max(\mathfrak{Z})$). By (9.4), $\mathfrak{f}_* \in \mathcal{S}_{(2k^2+1)k',k}(I_\nu)$, $\nu = 1, 2$, and hence, by Proposition 9.1 and Remark 9.2, we conclude that there exist two splines s_1, s_2 such that

$$s_1 \in S_{k',k} \cap \mathcal{M}^k[\mathfrak{f}_*, \mathfrak{p}](I_1), \quad s_2 \in S_{k',k} \cap \mathcal{M}^k[\mathfrak{p}, \mathfrak{f}_*](I_2),$$

and

$$\|\mathfrak{p} - s_\nu\|_{\mathbb{L}_p(I_\nu)} \le c_2 \|\mathfrak{p} - \mathfrak{f}_*\|_{\mathbb{L}_p(I_\nu)}. \tag{9.5}$$

Also note that $\mathfrak{f}_* \in \mathcal{S}_{k',k}[a, a_1]$ and $\mathfrak{f}_* \in \mathcal{S}_{k',k}[b_2, b]$, and define

$$s(x) = \begin{cases} \mathfrak{f}_*(x), & x \in [a, a_1] \cup [b_2, b], \\ s_1(x), & x \in [a_1, b_1], \\ \mathfrak{p}(x), & x \in [b_1, a_2], \\ s_2(x), & x \in [a_2, b_2]. \end{cases}$$

Then

$$s \in \mathcal{S}_{C(k),r} \cap \mathcal{M}^k[f](a, b), \quad C(k) \le 4k' + 1,$$

and, clearly, (9.3) is satisfied.

Case 3. $\#\mathfrak{Z} = \infty$. Clearly, the set of all limit points $\mathfrak{Z}^*$ is not empty in this case. Also, $\mathfrak{Z}^*$ is closed, and we now show that it has to be connected. This will imply that $\mathfrak{Z}^* = [c, d] \subset [a, b]$ (not excluding the possibility that $c = d$). Taking into account that $\mathfrak{f}_* - \mathfrak{p}$ is $(k-2)$ times continuously differentiable and has one-sided $(k-1)$st derivatives on $[a, b]$ (which is guaranteed by the assumption that $\mathfrak{f}_* \in \mathcal{M}^k[f]$), we apply Lemma 9.3 to conclude that, for every $x \in \mathfrak{Z}^*$, at least one of two relations takes place:

$$\mathfrak{f}_*^{(i)}(x\pm) = \mathfrak{p}^{(i)}(x), \quad i = 0, \dots, k-1.$$

Thus, if $\{c, d\} \subset \mathfrak{Z}^*$, then $\mathfrak{p} \in \mathcal{M}^k[\mathfrak{f}_*](c, d)$ so that the function

$$g_*(x) = \begin{cases} \mathfrak{f}_*(x), & x \in [a, b] \setminus [c, d], \\ \mathfrak{p}(x), & x \in [c, d], \end{cases}$$

is in $\mathcal{M}^k[\mathfrak{f}_*](a, b) \subset \mathcal{M}^k[f](a, b)$. Also, if $\mathfrak{f}_* \not\equiv \mathfrak{p}$ on $[c, d]$, then g_* approximates $\mathfrak{p}$ better (in the $\mathbb{L}_p$-metric) than $\mathfrak{f}_*$ on $[a, b]$ if $0 < p < \infty$ and not worse than $\mathfrak{f}_*$ if $p = \infty$. Therefore, we know (can assume) that $\mathfrak{f}_* \equiv \mathfrak{p}$ on $[c, d]$; hence $[c, d] \subset \mathfrak{Z}^*$.

Thus we can assume that $\mathfrak{Z}^* = [c, d]$ for some $[c, d] \subset [a, b]$. We also assume that $a < c \le d < b$, the cases in which $c = a$ or $d = b$ being analogous (and simpler).

Since $(a, c) \cap \mathfrak{Z}^* = \emptyset$, any closed subinterval of (a, c) contains finitely many points from $\mathfrak{Z}$.

Now, if $\#((a,c)\cap\mathfrak{Z})<2k^2+2$, (9.4) implies that $\mathfrak{f}_*\in\mathcal{S}_{(2k^2+2)k',k}[a,c]$, and we define the spline s_1 to be $\mathfrak{f}_*$ on $[a,c]$.

If, on the other hand, $\#((a,c)\cap\mathfrak{Z})\geq 2k^2+2$, then there exists $c'\in(a,c)$ such that $c'\in\mathfrak{Z}$, and the interval (a,c') contains exactly $2k^2+1$ points from $\mathfrak{Z}$. The same construction as in Case 2 allows us to obtain a k-monotone spline $\tilde{s}_1\in\mathcal{S}_{2k',k}(a,c')\cap\mathcal{M}^k[\mathfrak{f}_*,\mathfrak{p}]$ which "blends" $\mathfrak{f}_*$ with $\mathfrak{p}$ (in a k-monotone fashion) on (a,c') and approximates $\mathfrak{p}$ as well as $\mathfrak{f}_*$. Now we define s_1 by

$$s_1(x)=\begin{cases}\tilde{s}_1(x), & x\in[a,c'],\\ \mathfrak{p}(x), & x\in[c',c].\end{cases}$$

The same argument can now be used "at the right end" to yield a construction of $s_2\in\mathcal{S}_{(2k^2+2)k',k}[d,b]$ satisfying all conditions required.

Finally, we set

$$s(x)=\begin{cases}s_1(x), & x\in[a,c],\\ \mathfrak{p}(x), & x\in[c,d],\\ s_2(x), & x\in[d,b].\end{cases}$$

Then

$$s\in\mathcal{S}_{C(k),k}\cap\mathcal{M}^k[f][a,b],\quad C(k)\leq(4k^2+4)k'+1,$$

which completes the proof of Case 3 and of Proposition 4.2. □

REFERENCES

[1] R. K. Beatson, *Restricted range approximation by splines and variational inequalities*, SIAM J. Numer. Anal., 19 (1982), pp. 372–380.

[2] D. Braess, *Nonlinear Approximation Theory*, Springer-Verlag, Berlin, 1986.

[3] P. S. Bullen, *A criterion for n-convexity*, Pacific J. Math., 36 (1971), pp. 81–98.

[4] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Springer-Verlag, Berlin, 1993.

[5] Y.-K. Hu, *Convexity preserving approximation by free knot splines*, SIAM J. Math. Anal., 22 (1991), pp. 1183–1191.

[6] K. A. Kopotun, *Whitney theorem of interpolatory type for k-monotone functions*, Constr. Approx., 17 (2001), pp. 307–317.

[7] M. G. Krein and A. A. Nudel'man, *Problema Momentov Markova i Ekstremal'nye Zadachi*, Izdat. Nauka, Moscow, 1973.

[8] M. G. Krein and A. A. Nudel'man, *The Markov Moment Problem and Extremal Problems*, AMS, Providence, RI, 1977.

[9] D. Leviatan and A. Shadrin, *On monotone and convex approximation by splines with free knots*, Ann. Numer. Math., 4 (1997), pp. 415–434.

[10] G. G. Lorentz and K. L. Zeller, *Degree of approximation by monotone polynomials* I, J. Approx. Theory, 1 (1968), pp. 501–504.

[11] G. G. Lorentz and K. L. Zeller, *Degree of approximation by monotone polynomials* II, J. Approx. Theory, 2 (1969), pp. 265–269.

[12] A. Damas and M. Marano, *Property A and uniqueness in best approximation by n-convex functions*, in Trends in Approximation Theory, K. Kopotun, T. Lyche, and M. Neamtu, eds., Vanderbilt University Press, Nashville, TN, 2001, pp. 73–81.

[13] J. E. Pečarić, F. Proschan, and Y. L. Tong, *Convex Functions, Partial Orderings, and Statistical Applications*, Math. Sci. Engrg. 187, Academic Press, Boston, 1992.

[14] P. P. Petrov, *Shape preserving approximation by free knot splines*, East J. Approx., 2 (1996), pp. 41–48.

[15] P. P. Petrov, *Three-convex approximation by free knot splines in* $\mathbb{C}[0,1]$, Constr. Approx., 14 (1998), pp. 247–258.

[16] P. Petrushev, *Direct and converse theorems for spline and rational approximation and Besov spaces*, in Function Spaces and Applications, Lecture Notes in Math. 1302, M. Cwikel, J. Peetre, Y. Sagher, and H. Wallin, eds., Springer-Verlag, Berlin, 1988, pp. 363–377.

[17] P. Petrushev and V. Popov, *Rational Approximation of Real Functions*, Cambridge University Press, Cambridge, UK, 1987.
[18] A. W. Roberts and D. E. Varberg, *Convex Functions*, Academic Press, New York, 1973.
[19] V. A. Ubhaya, $\mathbb{L}_p$ *approximation from nonconvex subsets of special classes of functions*, J. Approx. Theory, 57 (1989), pp. 223–238.
[20] D. Zwick, *Existence of best n-convex approximations*, Proc. Amer. Math. Soc., 97 (1986), pp. 267–273.
[21] D. Zwick, *Best* $\mathbb{L}_1$*-approximation by generalized convex functions*, J. Approx. Theory, 59 (1989), pp. 116–123.

SIAM J. MATH. ANAL.
Vol. 34, No. 4, pp. 925–938

FORMATION OF δ-SHOCKS AND VACUUM STATES IN THE VANISHING PRESSURE LIMIT OF SOLUTIONS TO THE EULER EQUATIONS FOR ISENTROPIC FLUIDS*

GUI-QIANG CHEN† AND HAILIANG LIU‡

Abstract. The phenomena of concentration and cavitation and the formation of δ-shocks and vacuum states in solutions to the Euler equations for isentropic fluids are identified and analyzed as the pressure vanishes. It is shown that, as the pressure vanishes, any two-shock Riemann solution to the Euler equations for isentropic fluids tends to a δ-shock solution to the Euler equations for pressureless fluids, and the intermediate density between the two shocks tends to a weighted δ-measure that forms the δ-shock. By contrast, any two-rarefaction-wave Riemann solution of the Euler equations for isentropic fluids is shown to tend to a two-contact-discontinuity solution to the Euler equations for pressureless fluids, whose intermediate state between the two contact discontinuities is a vacuum state, even when the initial data stays away from the vacuum. Some numerical results exhibiting the formation process of δ-shocks are also presented.

Key words. concentration, cavitation, δ-shocks, vacuum states, Euler equations, vanishing pressure limit, transport equations, measure solutions, isentropic fluids, pressureless fluids, numerical simulations

AMS subject classifications. Primary, 35L65, 35B30, 76E19, 35Q35, 35L67; Secondary, 35B25, 65M06

PII. S0036141001399350

1. Introduction.

We are concerned with the phenomena of concentration and cavitation and the formation of δ-shocks and vacuum states in solutions to the Euler equations for compressible fluids as the pressure vanishes. In this paper, we consider the Euler equations of isentropic gas dynamics in Eulerian coordinates,

$$\partial_t \rho + \partial_x(\rho v) = 0, \tag{1.1}$$

$$\partial_t(\rho v) + \partial_x(\rho v^2 + P) = 0, \tag{1.2}$$

where ρ, P, and $m = \rho v$ represent the density, the scalar pressure, and the momentum, respectively; and ρ and m are in the physical region $\{(\rho, m) \,|\, \rho \geq 0, |m| \leq V_0\rho\}$ for some $V_0 > 0$. For $\rho > 0$, $v = m/\rho$ is the velocity with $|v| \leq V_0$. The scalar pressure P is a function of the density ρ and a small parameter $\epsilon > 0$ satisfying

$$\lim_{\epsilon \to 0} P(\rho, \epsilon) = 0.$$

For concreteness, we focus on the prototypical pressure function for polytropic gases:

$$P(\rho, \epsilon) = \epsilon p(\rho), \qquad p(\rho) = \rho^\gamma/\gamma, \quad \gamma > 1. \tag{1.3}$$

*Received by the editors December 8, 2001; accepted for publication (in revised form) August 9, 2002; published electronically March 5, 2003. The main observations and results in this paper were reported at the International Conference on Nonlinear Evolutionary Partial Differential Equations, Academia Sinica, China, 2001, and at the First Joint Meeting of the American Mathematical Society and the Société Mathématique de France, ENS, Lyon, France, 2001.

http://www.siam.org/journals/sima/34-4/39935.html

†Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208 (gqchen@math.northwestern.edu, http://www.math.northwestern.edu/∼gqchen). The research of this author was supported in part by the National Science Foundation under grants DMS-0204225, INT-9987378, and DMS-0204455.

‡Department of Mathematics, Iowa State University, Ames, IA 50011 (hliu@iastate.edu, http://www.math.iastate.edu/hliu). The research of this author was supported in part by the National Science Foundation under grant DMS-0107917.

System (1.1)–(1.3) is an example of hyperbolic systems of conservation laws with form

$$\partial_t u + \partial_x f(u, \epsilon) = 0, \tag{1.4}$$

with $u = (\rho, \rho v)$ and $f(u, \epsilon) = (\rho v, \rho v^2 + \epsilon p(\rho))$. Observe that system (1.1)–(1.3) with parameter $\epsilon > 0$ is generic in the sense that such a system can also be obtained under the scaling

$$(x, t) \longrightarrow (\alpha x, \alpha t), \qquad \rho \longrightarrow \alpha \rho,$$

with $\alpha = \epsilon^{-1/(\gamma-1)}$ from system (1.1)–(1.2) with $p = p(\rho)$.

In Chang, Chen, and Yang [4, 5, 6], a phenomenon of concentration in solutions of the two-dimensional Riemann problem was first observed numerically, which led to the occurrence of so-called smoothed δ-shocks for the Euler equations of gas dynamics when the Riemann data produces four initial contact discontinuities with different signs and the initial pressure data is close to zero. One of the main objectives of this paper is to show rigorously that the phenomenon of concentration in solutions, observed numerically in [4, 5, 6], for inviscid compressible fluid flow is fundamental and occurs not only in the multidimensional situations but also naturally in the one-dimensional case.

The limit system as $\epsilon \to 0$ formally becomes the transport equations

$$\partial_t \rho + \partial_x (\rho v) = 0, \tag{1.5}$$

$$\partial_t (\rho v) + \partial_x (\rho v^2) = 0, \tag{1.6}$$

which are also called the one-dimensional system of pressureless Euler equations, modeling the motion of free particles which stick under collision (see [3, 11, 30]).

The transport equations (1.5)–(1.6) have been analyzed extensively since 1994; for example, see [1, 2, 3, 11, 12, 13, 17, 18, 19, 20, 22, 28, 29] and the references cited therein. In particular, the existence of measure solutions of the Riemann problem was first presented in Bouchut [1], and a connection of (1.5)–(1.6) with adhesion particle dynamics and the behavior of global weak solutions with random initial data were discussed in E, Rykov, and Sinai [11]. Also see [14, 15, 16, 21, 26, 27] for related equations and results. It has been shown that, for the transport equations (1.5)–(1.6), δ-shocks and vacuum states do occur in the Riemann solutions. Since the two eigenvalues of the transport equations coincide, the occurrence of δ-shocks and vacuum states as $t > 0$ can be regarded as a result of resonance between the two characteristic fields.

In this paper, we rigorously analyze the phenomena of concentration and cavitation and the formation of δ-shocks and vacuum states in solutions to the Euler equations for isentropic fluids as the pressure vanishes. The vanishing pressure limit can be regarded as a singular flux-function limit for hyperbolic systems of conservation laws (1.4). We show that such phenomena occur naturally in the one-dimensional case as the pressure vanishes: any two-shock Riemann solution to the Euler equations for isentropic fluids tends to a δ-shock solution to the Euler equations for pressureless fluids, and the intermediate density between the two shocks tends to a weighted δ-measure that forms a δ-shock; by contrast, any two-rarefaction-wave Riemann solution to the Euler equations for isentropic fluids tends to a two-contact-discontinuity solution to the Euler equations for pressureless fluids, whose intermediate state between the two contact discontinuities is a vacuum state even when the initial data

stays away from the vacuum. These results show that the δ-shocks for the transport equations result from a phenomenon of concentration, while the vacuum states result from a phenomenon of cavitation in the process of the vanishing pressure limit; both are fundamental and physical in fluid dynamics.

From the point of view of hyperbolic conservation laws, since the limit system loses hyperbolicity, the phenomena of concentration and cavitation in the process of the vanishing pressure limit can be regarded as phenomena of resonance between the two characteristic fields. These phenomena show that the flux-function limit can be very singular: the limit functions of solutions are no longer in the spaces of functions BV or L^∞; and the space of Radon measures, for which the divergences of certain entropy and entropy flux fields are also Radon measures, is a natural space in order to deal with such a limit in general. In this regard, a theory of divergence-measure fields has been established in Chen and Frid [7, 8, 9].

The organization of this paper is as follows. In section 2, we discuss the δ-shocks and vacuum states for the transport equations (1.5)–(1.6) and examine the dependence of the Riemann solutions on the parameter $\epsilon > 0$ for the Euler equations (1.1)–(1.3). In section 3, we analyze the formation of δ-shocks in the Riemann solutions to the Euler equations (1.1)–(1.3) as the pressure vanishes. In section 4, we analyze the formation of vacuum states in the Riemann solutions to (1.1)–(1.3), even when the initial data stays away from the vacuum, as the pressure decreases. In section 5, we present some representative numerical results, produced by using the higher order essentially nonoscillatory (ENO) scheme in [23, 24], to examine the phenomenon of concentration and the formation process of δ-shocks in the level of the Euler dynamics (1.1)–(1.3) as the pressure decreases.

2. δ-shocks, vacuum states, and Riemann solutions. In this section, we first discuss δ-shocks and vacuum states in the Riemann solutions to the transport equations (1.5)–(1.6), and then we examine the dependence of the Riemann solutions on the parameter $\epsilon > 0$ to the Euler equations (1.1)–(1.3).

2.1. δ-shocks and vacuum states for the transport equations. Consider the Riemann problem of the transport equations (1.5)–(1.6) with Riemann initial data

$$(\rho, v)(x, 0) = (\rho_\pm, v_\pm), \qquad \pm x > 0, \tag{2.1}$$

with $\rho_\pm > 0$. Since the equations and the Riemann data are invariant under uniform stretching of coordinates

$$(x, t) \to (\beta x, \beta t), \qquad \beta \text{ constant},$$

we consider the self-similar solutions of (1.5), (1.6), and (2.1):

$$(\rho, v)(x, t) = (\rho, v)(\xi), \quad \xi = x/t.$$

Then the Riemann problem is reduced to a boundary value problem for ordinary differential equations:

$$\begin{aligned} &-\xi\rho_\xi + (\rho v)_\xi = 0,\\ &-\xi(\rho v)_\xi + (\rho v^2)_\xi = 0,\\ &(\rho, v)(\pm\infty) = (\rho_\pm, v_\pm). \end{aligned}$$

As shown in [22], in the case in which $v_- < v_+$, we can obtain a solution $(\rho, v)(\xi)$ that consists of two contact discontinuities and a vacuum state which are uniquely determined by the Riemann data $(\rho_\pm, v_\pm)$. That is,

$$(\rho, v)(\xi) = \begin{cases} (\rho_-, v_-), & -\infty < \xi \le v_-, \\ (0, \xi), & v_- \le \xi \le v_+, \\ (\rho_+, v_+), & v_+ \le \xi < \infty. \end{cases}$$

In the case in which $v_- > v_+$, a key observation in [22] is that the singularity cannot be a jump with finite amplitude; that is, there is no solution which is piecewise smooth and bounded. Hence a solution containing a weighted δ-measure (i.e., δ-shock) supported on a line was constructed in order to establish the existence in a space of measures from the mathematical point of view (see also [26, 27]).

To define the measure solutions, the weighted δ-measure $w(t)\delta_S$ supported on a smooth curve $S = \{(x(s), t(s)) : a < s < b\}$ can be defined by

$$\langle w(\cdot)\delta_S, \psi(\cdot,\cdot)\rangle = \int_a^b w(t(s))\psi(x(s), t(s))\sqrt{x'(s)^2 + t'(s)^2}\, ds$$

for any $\psi \in C_0^\infty((-\infty, \infty) \times [0, \infty))$.

With this definition, a family of δ-measure solutions with parameter σ in the case in which $v_- > v_+$ can be obtained as

$$\rho(x,t) = \rho_0(x,t) + w(t)\delta_S, \qquad v(x,t) = v_0(x,t),$$

where $S = \{(\sigma t, t) : 0 \le t < \infty\}$,

$$\rho_0(x,t) = \rho_- + [\rho]\,\chi(x - \sigma t), \;\; v_0(x,t) = v_- + [v]\,\chi(x - \sigma t), \;\; w(t) = \frac{t}{1+\sigma^2}(\sigma[\rho] - [\rho v]),$$

in which $[h] := h_+ - h_-$ denotes the jump of function h across the discontinuity, and $\chi(x)$ is the characteristic (or indicator) function that is 0 when $x < 0$ and 1 when $x > 0$.

It was shown in [22] that the δ-measure solutions (ρ, v) constructed above satisfy

$$\langle \rho, \phi_t\rangle + \langle \rho v, \phi_x\rangle = 0, \tag{2.2}$$

$$\langle \rho v, \phi_t\rangle + \langle \rho v^2, \phi_x\rangle = 0 \tag{2.3}$$

for any $\phi \in C_0^\infty((-\infty, \infty) \times (0, \infty))$, where

$$\langle \rho, \phi\rangle = \int_0^\infty \int_{-\infty}^\infty \rho_0 \phi \, dx dt + \langle w\delta_S, \phi\rangle$$

and

$$\langle \rho v, \phi\rangle = \int_0^\infty \int_{-\infty}^\infty \rho_0 v_0 \phi \, dx dt + \langle \sigma w \delta_S, \phi\rangle.$$

A unique solution can be singled out by the so-called δ-Rankine–Hugoniot condition

$$\sigma = \frac{\sqrt{\rho_+}v_+ + \sqrt{\rho_-}v_-}{\sqrt{\rho_+} + \sqrt{\rho_-}} \tag{2.4}$$

that satisfies the δ-entropy condition

$$v_+ < \sigma < v_-. \tag{2.5}$$

The entropy condition (2.5) means that, in the (x,t)-plane, all the characteristic lines on either side of a δ-shock run into the line of δ-shock, which implies that a δ-shock is an overcompressive shock.

2.2. Riemann solutions to the Euler equations for isentropic fluids. The eigenvalues of system (1.1)–(1.3) are

$$\lambda_1 = v - c(\rho,\epsilon), \quad \lambda_2 = v + c(\rho,\epsilon) \qquad \text{for } \rho > 0$$

with

$$c(\rho,\epsilon) = \sqrt{\epsilon\, p'(\rho)} = \sqrt{\epsilon}\,\rho^{\theta}, \qquad \theta = \frac{\gamma-1}{2}.$$

The Riemann invariants are

$$w = v + \int_0^{\rho} \frac{\sqrt{\epsilon p'(s)}}{s} ds, \qquad z = v - \int_0^{\rho} \frac{\sqrt{\epsilon p'(s)}}{s} ds.$$

Then the Riemann solutions, which are functions of $\xi = x/t$, are solutions of

$$-\xi\rho_\xi + (\rho v)_\xi = 0, \tag{2.6}$$

$$-\xi(\rho v)_\xi + (\rho v^2 + \epsilon\, p(\rho))_\xi = 0, \tag{2.7}$$

$$(\rho, v)(\pm\infty) = (\rho_\pm, v_\pm). \tag{2.8}$$

Shock curves. The Rankine–Hugoniot conditions for discontinuous solutions of (1.1)–(1.3) are

$$-\sigma[\rho] + [\rho v] = 0, \qquad -\sigma[\rho v] + [\rho v^2 + \epsilon\, p(\rho)] = 0.$$

The Lax entropy inequalities imply

$$\rho_+ > \rho_- \ \text{(one-shock)}, \qquad \rho_+ < \rho_- \ \text{(two-shock)}.$$

Then, given a state $u_- = (\rho_-, \rho_- v_-)$, the shock curves in the phase plane, which are the sets of states that can be connected on the right by a one-shock or a two-shock, are the following.

One-shock curve $S_1(u_-)$:

$$v - v_- = -\sqrt{\frac{\epsilon\,(p(\rho) - p(\rho_-))}{\rho_-\rho(\rho-\rho_-)}}(\rho-\rho_-), \qquad \rho > \rho_-.$$

Two-shock curve $S_2(u_-)$:

$$v - v_- = -\sqrt{\frac{\epsilon\,(p(\rho) - p(\rho_-))}{\rho_-\rho(\rho-\rho_-)}}(\rho-\rho_-), \qquad \rho < \rho_-.$$

Then the shock curves are concave or convex, respectively, with respect to the point $u_- = (\rho_-, \rho_- v_-)$ in the $\rho - m$ plane with $m = \rho v$; that is, the quotient $\frac{m-m_-}{\rho-\rho_-}$ as a function of ρ is monotone.

We now turn to analyzing the Riemann solutions that consist of rarefaction waves and constant states. There are also two families of rarefaction waves, corresponding to characteristic fields λ_1 and λ_2, respectively.

Rarefaction wave curves. A rarefaction wave is a continuous solution of (2.6)–(2.8) of the form $(\rho, \rho v)(\xi)$, $\xi = x/t$, satisfying

$$\xi = v \mp \sqrt{\epsilon\, p'(\rho)}, \qquad -\xi\rho_\xi + (\rho v)_\xi = 0.$$

Then, given a state $u_- = (\rho_-, \rho_- v_-)$, the rarefaction-wave curves in the phase plane, which are the sets of states that can be connected on the right by a one-rarefaction or two-rarefaction wave, are the following.

One-rarefaction wave curve $R_1(u_-)$:

$$v - v_- = -\int_{\rho_-}^{\rho} \frac{\sqrt{\epsilon\, p'(s)}}{s} ds, \qquad \rho < \rho_-.$$

Two-rarefaction wave curve $R_2(u_-)$:

$$v - v_- = \int_{\rho_-}^{\rho} \frac{\sqrt{\epsilon\, p'(s)}}{s} ds, \qquad \rho > \rho_-.$$

The rarefaction wave curves are concave or convex, respectively, in the $\rho - m$ plane.

Given a left state $u_- = (\rho_-, \rho_- v_-)$, the set of states that can be connected on the right by a shock or a rarefaction wave in the phase plane consists of the one-shock curve $S_1(u_-)$, the one-rarefaction curve $R_1(u_-)$, the two-shock curve $S_2(u_-)$, and the two-rarefaction curve $R_2(u_-)$. These curves divide the phase plane into four regions $S_2S_1(u_-)$, $S_2R_1(u_-)$, $R_2S_1(u_-)$, and $R_2R_1(u_-)$; any right state of the Riemann data staying in one of them yields a unique global Riemann solution $R(x/t)$, which contains a one-shock (or a one-rarefaction wave) and/or a two-shock (or a two-rarefaction wave) satisfying

$$w(R(x/t)) \le w(u_+), \;\; z(R(x/t)) \ge z(u_-), \;\; w(R(x/t)) - z(R(x/t)) \ge 0.$$

In particular, when $u_+ \in S_2S_1(u_-)$, $R(x/t)$ contains a one-shock, a two-shock, and a nonvacuum intermediate constant state; and, when $u_+ \in R_2R_1(u_-)$, $R(x/t)$ contains a one-rarefaction wave, a two-rarefaction wave, and an intermediate constant state that may be a vacuum state. Since the other two regions $S_2R_1(u_-)$ and $R_2S_1(u_-)$ have empty interiors when $\epsilon \to 0$, it suffices to analyze the limit process for the two cases $u_+ \in S_2S_1(u_-)$ (in section 3) and $u_+ \in R_2R_1(u_-)$ (in section 4). For more details about Riemann solutions, see [10, 25].

3. Formation of δ-shocks. In this section, we study the formation of δ-shocks in the Riemann solutions to the Euler equations for isentropic fluids in the case $u_+ \in S_2S_1(u_-)$ with $v_- > v_+$ and $\rho_\pm > 0$ as the pressure vanishes.

3.1. Limiting behavior of the Riemann solutions as $\epsilon \to 0$. For fixed $\epsilon > 0$, let $u_*^\epsilon := (\rho_*^\epsilon, \rho_*^\epsilon v_*^\epsilon)$ be the intermediate state in the sense that u_- and u_*^ϵ are connected by one-shock S_1 with speed σ_1^ϵ and that u_*^ϵ and u_+ are connected by two-shock S_2 with speed σ_2^ϵ. Then $(\rho_*^\epsilon, v_*^\epsilon)$ are determined by

$$v_*^\epsilon - v_- = -\sqrt{\frac{\epsilon\,(p(\rho_*^\epsilon) - p(\rho_-))}{\rho_-\rho_*^\epsilon(\rho_*^\epsilon - \rho_-)}}(\rho_*^\epsilon - \rho_-), \qquad \rho_*^\epsilon > \rho_-, \tag{3.1}$$

and

$$v_+ - v_*^\epsilon = -\sqrt{\frac{\epsilon\,(p(\rho_+) - p(\rho_*^\epsilon))}{\rho_+\rho_*^\epsilon(\rho_+ - \rho_*^\epsilon)}}(\rho_+ - \rho_*^\epsilon), \qquad \rho_*^\epsilon > \rho_+. \tag{3.2}$$

Define $g(s,\tau)=\sqrt{(\frac{1}{s}-\frac{1}{\tau})(p(\tau)-p(s))}$ for $s,\tau>0$. Thus a combination of the jump conditions (3.1) and (3.2) gives

$$v_- - v_+ = \sqrt{\epsilon}\,(g(\rho_*^\epsilon,\rho_-)+g(\rho_*^\epsilon,\rho_+))>0. \tag{3.3}$$

Then one must have $\lim_{\epsilon\to 0} g(\rho_*^\epsilon,\rho_\pm)=\infty$, which yields $\lim_{\epsilon\to 0}\rho_*^\epsilon=\infty$. Letting $\epsilon\to 0$ in (3.3) yields

$$\lim_{\epsilon\to 0}\epsilon(\rho_*^\epsilon)^\gamma=\frac{\sqrt{\rho_-\rho_+}}{\sqrt{\rho_-}+\sqrt{\rho_+}}(v_- - v_+).$$

Therefore, we have the following lemma.

LEMMA 3.1. $\lim_{\epsilon\to 0}\epsilon^{1/\gamma}\rho_*^\epsilon=(\frac{\sqrt{\rho_-\rho_+}}{\sqrt{\rho_-}+\sqrt{\rho_+}}(v_- - v_+))^{1/\gamma}$.

LEMMA 3.2. *Set* $\sigma=\frac{\sqrt{\rho_-}v_-+\sqrt{\rho_+}v_+}{\sqrt{\rho_-}+\sqrt{\rho_+}}\in(v_+,v_-)$. *Then*

$$\lim_{\epsilon\to 0} v_*^\epsilon=\lim_{\epsilon\to 0}\sigma_1^\epsilon=\lim_{\epsilon\to 0}\sigma_2^\epsilon=\sigma$$

and

$$\lim_{\epsilon\to 0}\rho_*^\epsilon(\sigma_2^\epsilon-\sigma_1^\epsilon)=\sigma[\rho]-[\rho v].$$

Proof. First, Lemma 3.1 and (3.1)–(3.2) immediately imply that

$$\lim_{\epsilon\to 0} v_*^\epsilon=\sigma.$$

On the other hand, using the Lax entropy inequalities for the shocks, we have

$$v_*^\epsilon-\sqrt{\epsilon}\,(\rho_*^\epsilon)^\theta<\sigma_1^\epsilon<\min(v_*^\epsilon+\sqrt{\epsilon}\,(\rho_*^\epsilon)^\theta, v_- -\sqrt{\epsilon}\,(\rho_-)^\theta) \tag{3.4}$$

and

$$\max(v_*^\epsilon-\sqrt{\epsilon}\,(\rho_*^\epsilon)^\theta, v_+ +\sqrt{\epsilon}\,(\rho_+)^\theta)<\sigma_2^\epsilon<v_*^\epsilon+\sqrt{\epsilon}\,(\rho_*^\epsilon)^\theta. \tag{3.5}$$

Noting that $\sqrt{\epsilon}\,(\rho_*^\epsilon)^\theta=\epsilon^{\frac{1}{2\gamma}}(\epsilon^{1/\gamma}\rho_*^\epsilon)^\theta$, we can see from Lemma 3.1 that, for $\gamma>1$,

$$\lim_{\epsilon\to 0}\sqrt{\epsilon}\,(\rho_*^\epsilon)^\theta=0. \tag{3.6}$$

Then (3.4)–(3.6) yield

$$\lim_{\epsilon\to 0} v_*^\epsilon=\lim_{\epsilon\to 0}\sigma_1^\epsilon=\lim_{\epsilon\to 0}\sigma_2^\epsilon=\sigma. \tag{3.7}$$

The Rankine–Hugoniot conditions for (1.1) on the shocks and the results of (3.7) yield

$$\rho_*^\epsilon(\sigma_2^\epsilon-\sigma_1^\epsilon)=\sigma_2^\epsilon\,\rho_+-\sigma_1^\epsilon\,\rho_--[\rho v]\to\sigma[\rho]-[\rho v]\qquad\text{as}\quad\epsilon\to 0.$$

This completes the proof of Lemma 3.2. ∎

Remark 3.1. The quantity σ that is the limit of v_*^ϵ, σ_1^ϵ, and σ_2^ϵ uniquely determines the δ-shock solution of (1.5)–(1.6) as the limit of the Riemann solutions when $\epsilon\to 0$ and is consistent with the δ-Rankine–Hugoniot condition (2.4) and the δ-entropy condition (2.5), as proposed for the Riemann solutions for pressureless Euler equations.

3.2. Weighted δ-shocks. We now show the following theorem characterizing the vanishing pressure limit in the case in which $v_- > v_+$.

THEOREM 3.1. *Let $v_- > v_+$. For each fixed $\epsilon > 0$, assume that $(\rho^\epsilon, m^\epsilon) = (\rho^\epsilon, \rho^\epsilon v^\epsilon)$ is a two-shock solution of (1.1)–(1.3) with Riemann data $u_\pm = (\rho_\pm, \rho_\pm v_\pm)$, constructed in section 2.2. Then, when $\epsilon \to 0$, ρ^ϵ and m^ϵ converge in the sense of distributions, and the limit functions ρ and m are the sums of a step function and a δ-measure with weights*

$$\frac{t}{\sqrt{1+\sigma^2}}(\sigma[\rho] - [\rho v]) \quad \text{and} \quad \frac{t}{\sqrt{1+\sigma^2}}(\sigma[\rho v] - [\rho v^2]),$$

respectively, which form a δ-shock solution of (1.5)–(1.6) with the same Riemann data $u_\pm$.

Proof. 1. Set $\xi = x/t$. Then, for each fixed $\epsilon > 0$, the Riemann solution can be written as

$$\rho^\epsilon(\xi) = \begin{cases} \rho_- & \text{for } \xi < \sigma_1^\epsilon, \\ \rho_*^\epsilon(\xi) & \text{for } \sigma_1^\epsilon < \xi < \sigma_2^\epsilon, \\ \rho_+ & \text{for } \xi > \sigma_2^\epsilon \end{cases}$$

and

$$v^\epsilon(\xi) = \begin{cases} v_- & \text{for } \xi < \sigma_1^\epsilon, \\ v_*^\epsilon(\xi) & \text{for } \sigma_1^\epsilon < \xi < \sigma_2^\epsilon, \\ v_+ & \text{for } \xi > \sigma_2^\epsilon, \end{cases}$$

satisfying the following weak formulations: For any $\psi \in C_0^1(-\infty, \infty)$,

$$-\int_{-\infty}^{\infty} (v^\epsilon(\xi) - \xi)\rho^\epsilon(\xi)\,\psi'(\xi)d\xi + \int_{-\infty}^{\infty} \rho^\epsilon(\xi)\,\psi(\xi)\,d\xi = 0, \tag{3.8}$$

and

$$\begin{aligned} -\int_{-\infty}^{\infty} (v^\epsilon(\xi) - \xi)\rho^\epsilon(\xi)v^\epsilon(\xi)\,\psi'(\xi)d\xi + \int_{-\infty}^{\infty} \rho^\epsilon(\xi)v^\epsilon(\xi)\,\psi(\xi)d\xi \\ = \epsilon \int_{-\infty}^{\infty} p(\rho^\epsilon(\xi))\,\psi'(\xi)d\xi. \end{aligned} \tag{3.9}$$

2. The first integral in (3.8) can be decomposed into

$$-\left\{\int_{-\infty}^{\sigma_1^\epsilon} + \int_{\sigma_1^\epsilon}^{\sigma_2^\epsilon} + \int_{\sigma_2^\epsilon}^{\infty}\right\} (v^\epsilon(\xi) - \xi)\rho^\epsilon(\xi)\psi'(\xi)d\xi. \tag{3.10}$$

The sum of the first and last term of (3.10) is

$$\begin{aligned} &-\int_{-\infty}^{\sigma_1^\epsilon} (v_- - \xi)\rho_-\psi'(\xi)d\xi - \int_{\sigma_2^\epsilon}^{\infty} (v_+ - \xi)\rho_+\psi'(\xi)d\xi \\ &= -\rho_- v_-\psi(\sigma_1^\epsilon) + \rho_+ v_+\psi(\sigma_2^\epsilon) + \rho_-\sigma_1^\epsilon\psi(\sigma_1^\epsilon) - \rho_+\sigma_2^\epsilon\psi(\sigma_2^\epsilon) \\ &\quad - \int_{-\infty}^{\sigma_1^\epsilon} \rho_-\psi(\xi)d\xi - \int_{\sigma_2^\epsilon}^{\infty} \rho_+\psi(\xi)d\xi, \end{aligned}$$

which converges as $\epsilon \to 0$ to

$$([\rho v] - \sigma[\rho])\,\psi(\sigma) - \int_{-\infty}^{\infty} \rho_0(\xi - \sigma)\,\psi(\xi)d\xi$$

with

$$\rho_0(\xi) = \rho_- + [\rho]\,\chi(\xi),$$

where $\chi(\xi)$ is the characteristic function.

For the second term of (3.10),

$$\begin{aligned}
&\lim_{\epsilon\to 0}\int_{\sigma_1^\epsilon}^{\sigma_2^\epsilon} (v^\epsilon(\xi) - \xi)\rho^\epsilon(\xi)\psi'(\xi)d\xi \\
&\quad = \rho_*^\epsilon(\sigma_2^\epsilon - \sigma_1^\epsilon)\left\{ v_*^\epsilon \frac{\psi(\sigma_2^\epsilon) - \psi(\sigma_1^\epsilon)}{\sigma_2^\epsilon - \sigma_1^\epsilon} - \frac{\sigma_2^\epsilon\psi(\sigma_2^\epsilon) - \sigma_1^\epsilon\psi(\sigma_1^\epsilon)}{\sigma_2^\epsilon - \sigma_1^\epsilon} + \frac{1}{\sigma_2^\epsilon - \sigma_1^\epsilon}\int_{\sigma_1^\epsilon}^{\sigma_2^\epsilon} \psi(\xi)d\xi \right\},
\end{aligned}$$

which converges as $\epsilon \to 0$ to

$$([\rho v] - \sigma[\rho])\{-\sigma\psi'(\sigma) + \sigma\psi'(\sigma) + \psi(\sigma) - \psi(\sigma)\} = 0$$

since $\psi \in C_0^1(-\infty, \infty)$, $\lim_{\epsilon\to 0} v_*^\epsilon = \sigma$, and $\lim_{\epsilon\to 0} \sigma_j^\epsilon = \sigma$ for $j = 1, 2$.

Then the integral identity (3.8) yields

$$\lim_{\epsilon\to 0}\int_{-\infty}^{\infty} (\rho^\epsilon(\xi) - \rho_0(\xi - \sigma))\,\psi(\xi)d\xi = (\sigma[\rho] - [\rho v])\,\psi(\sigma)$$

for any function $\psi \in C_0^\infty(-\infty, \infty)$.

3. We now turn to justifying the limit of momentum $m^\epsilon = \rho^\epsilon v^\epsilon$ using the weak formulation (3.9). As done previously, we can obtain the limit for the first term on the left of (3.9) as

$$\begin{aligned}
&-\lim_{\epsilon\to 0}\int_{-\infty}^{\infty} (v^\epsilon(\xi) - \xi)\,\rho^\epsilon(\xi)v^\epsilon(\xi)\psi'(\xi)d\xi \\
&= \psi(\sigma)\,([\rho v^2] - \sigma[\rho v]) - \int_{-\infty}^{\sigma} \rho_- v_-\,\psi(\xi)d\xi - \int_{\sigma}^{\infty} \rho_+ v_+\,\psi(\xi)d\xi.
\end{aligned}$$

The term on the right of (3.9) equals

$$\epsilon\int_{-\infty}^{\infty} p(\rho^\epsilon)\,\psi'(\xi)d\xi = \epsilon\left\{\int_{-\infty}^{\sigma_1^\epsilon} + \int_{\sigma_1^\epsilon}^{\sigma_2^\epsilon} + \int_{\sigma_2^\epsilon}^{\infty}\right\} p(\rho^\epsilon)\,\psi'(\xi)d\xi,$$

which converges to

$$\begin{aligned}
&\epsilon\,\{p(\rho_-)\psi(\sigma_1^\epsilon) + p(\rho_*^\epsilon)(\psi(\sigma_2^\epsilon) - \psi(\sigma_1^\epsilon)) - p(\rho_+)\psi(\sigma_2^\epsilon)\} \\
&\quad = o(\epsilon) + \epsilon\,p(\rho_*^\epsilon)(\psi(\sigma_2^\epsilon) - \psi(\sigma_1^\epsilon)) \to 0 \qquad \text{as } \epsilon \to 0,
\end{aligned}$$

where we used the fact that $\epsilon p(\rho_*^\epsilon)$ is bounded and $\lim_{\epsilon\to 0} \sigma_j^\epsilon = \sigma$ for $j = 1, 2$.

Returning to the weak formulation (3.9), one has

$$\lim_{\epsilon\to 0}\int_{-\infty}^{\infty} ((\rho^\epsilon v^\epsilon)(\xi) - (\rho_0 v_0)(\xi - \sigma))\,\psi(\xi)d\xi = \psi(\sigma)(\sigma[\rho v] - [\rho v^2]).$$

4. Finally, we are in a position to study the limits of ρ^ϵ and $m^\epsilon = \rho^\epsilon v^\epsilon$ by tracking the time-dependence of the weights of the δ-measures as $\epsilon \to 0$.

Let $\phi(x,t) \in C_0^\infty((-\infty,\infty)\times[0,\infty))$ be a smooth test function, and let $\tilde{\phi}(\xi,t) := \phi(\xi t, t)$. Then we have

$$\lim_{\epsilon\to 0}\int_0^\infty\int_{-\infty}^\infty \rho^\epsilon(x/t)\phi(x,t)dxdt = \lim_{\epsilon\to 0}\int_0^\infty t\left(\int_{-\infty}^\infty \rho^\epsilon(\xi)\tilde{\phi}(\xi,t)d\xi\right)dt.$$

On the other hand, we have

$$\begin{aligned}\lim_{\epsilon\to 0}\int_{-\infty}^\infty \rho^\epsilon(\xi)\tilde{\phi}(\xi,t)d\xi &= \int_{-\infty}^\infty \rho_0(\xi-\sigma)\tilde{\phi}(\xi,t)d\xi + (\sigma[\rho]-[\rho v])\tilde{\phi}(\sigma,t)\\ &= t^{-1}\int_{-\infty}^\infty \rho_0(x-\sigma t)\phi(x,t)dx + (\sigma[\rho]-[\rho v])\phi(\sigma t,t).\end{aligned}$$

Combining the two relations above yields

$$\begin{aligned}&\lim_{\epsilon\to 0}\int_0^\infty\int_{-\infty}^\infty \rho^\epsilon(x/t)\phi(x,t)dxdt\\ &\quad= \int_0^\infty\int_{-\infty}^\infty \rho_0(x-\sigma t)\phi(x,t)dxdt + \int_0^\infty t([\rho v]-\sigma[\rho])\phi(\sigma t,t)dt.\end{aligned}$$

The last term, by the definition, equals

$$\langle w_1(\cdot)\delta_S, \phi(\cdot,\cdot)\rangle$$

with

$$w_1(t) = \frac{t}{\sqrt{1+\sigma^2}}(\sigma[\rho]-[\rho v]).$$

Similarly, we can show that

$$\begin{aligned}&\lim_{\epsilon\to 0}\int_0^\infty\int_{-\infty}^\infty m^\epsilon(x/t)\phi(x,t)dxdt\\ &\quad= \int_0^\infty\int_{-\infty}^\infty (\rho_0 v_0)(x-\sigma t)\phi(x,t)dxdt + \langle w_2(\cdot)\delta_S, \phi(\cdot,\cdot)\rangle\end{aligned}$$

with

$$w_2(t) = \frac{t}{\sqrt{1+\sigma^2}}(\sigma[\rho v]-[\rho v^2]).$$

This completes the proof of Theorem 3.1. □

4. Formation of vacuum states. In this section, we show the formation of vacuum states in the Riemann solutions of (1.1)–(1.3) in the case in which $u_+ \in R_2R_1(u_-)$ with $v_- < v_+$ and $\rho_\pm > 0$ as the pressure decreases.

As stated previously, on the rarefaction waves, the solution satisfies

$$\xi = x/t = v^\epsilon \pm \sqrt{\epsilon\, p'(\rho^\epsilon)} \tag{4.1}$$

for each fixed $\epsilon > 0$. More precisely, we have that, on the one-rarefaction wave,

$$\xi = v^\epsilon - \sqrt{\epsilon\, p'(\rho^\epsilon)}, \qquad v_- - \sqrt{\epsilon\, p'(\rho_-)} < \xi < v_*^\epsilon - \sqrt{\epsilon\, p'(\rho_*^\epsilon)}, \qquad \rho_- > \rho_*^\epsilon,$$

and, on the two-rarefaction wave,

$$\xi = v^\epsilon + \sqrt{\epsilon\, p'(\rho^\epsilon)}, \qquad v_*^\epsilon + \sqrt{\epsilon\, p'(\rho_*^\epsilon)} < \xi < v_+ + \sqrt{\epsilon\, p'(\rho_+)}, \qquad \rho_*^\epsilon < \rho_+,$$

where $(\rho_*^\epsilon, \rho_*^\epsilon v_*^\epsilon)$ is the intermediate state in the Riemann solutions. Since $(\rho_*^\epsilon, \rho_*^\epsilon v_*^\epsilon)$ is on the curve $R_1(u_-)$, we have

$$v_*^\epsilon = v_- - \int_{\rho_-}^{\rho_*^\epsilon} \frac{\sqrt{\epsilon\, p'(s)}}{s} ds \le v_- + \int_0^{\rho_-} \frac{\sqrt{\epsilon\, p'(s)}}{s} ds = v_- + \sqrt{\epsilon}\frac{\rho_-^\theta}{\theta} \equiv A^\epsilon.$$

When $v_- < v_+ < A^\epsilon$, that is,

$$\epsilon > \left(\frac{\theta(v_+ - v_-)}{\rho_-^\theta}\right)^2 \equiv \epsilon_0(u_-, u_+), \tag{4.2}$$

there is no vacuum in the solution. This implies that, for a fluid with strong pressure, no vacuum occurs in the solution generically.

However, when ϵ decreases so that $\epsilon < \epsilon_0(u_-, u_+)$, then $A^\epsilon < v_+$, and the intermediate state $(\rho_*^\epsilon, \rho_*^\epsilon v_*^\epsilon)$ becomes a vacuum state with

$$(\rho_*^\epsilon, v_*^\epsilon)(\xi) = (0, \xi), \qquad v_1^\epsilon \le \xi \le v_2^\epsilon,$$

where

$$v_1^\epsilon = v_- + \int_0^{\rho_-} \frac{\sqrt{\epsilon p'(s)}}{s} ds, \quad v_2^\epsilon = v_+ - \int_0^{\rho_+} \frac{\sqrt{\epsilon p'(s)}}{s} ds.$$

The uniform boundedness of $\rho^\epsilon(\xi)$ with respect to ϵ in this case leads to

$$\lim_{\epsilon\to 0} v_1^\epsilon = v_-, \qquad \lim_{\epsilon\to 0} v_2^\epsilon = v_+,$$

and

$$\lim_{\epsilon\to 0} v^\epsilon(\xi) = \xi \qquad \text{for } \xi \in (v_-, v_+).$$

In summary, the limit function (ρ, v) in this case is

$$(\rho, v)(\xi) = \begin{cases} (\rho_-, v_-), & -\infty < \xi \le v_-, \\ (0, \xi), & v_- \le \xi \le v_+, \\ (\rho_+, v_+), & v_+ \le \xi < \infty, \end{cases}$$

which is a solution to the transport equations (1.5)–(1.6) containing a vacuum state that fills up the region formed by the two contact discontinuities $\xi = x/t = v_\pm$.

We can clearly see from the analysis above that, when ϵ decreases, the left boundary of the one-rarefaction wave and the right boundary of the two-rarefaction wave are fixed, the right boundary of the one-rarefaction wave becomes closer to the left boundary of the one-rarefaction wave, and the left boundary of the two-rarefaction wave becomes closer to the right boundary of the two-rarefaction wave; while the state between the right boundary of the one-rarefaction wave and the left boundary of the two-rarefaction wave in the Riemann solution is a vacuum state; and, in the limit, the left boundary of the one-rarefaction wave and the right boundary of the two-rarefaction wave become two contact discontinuities of the transport equations (1.5)–(1.6), and the vacuum state fills up the region between the two contact discontinuities.

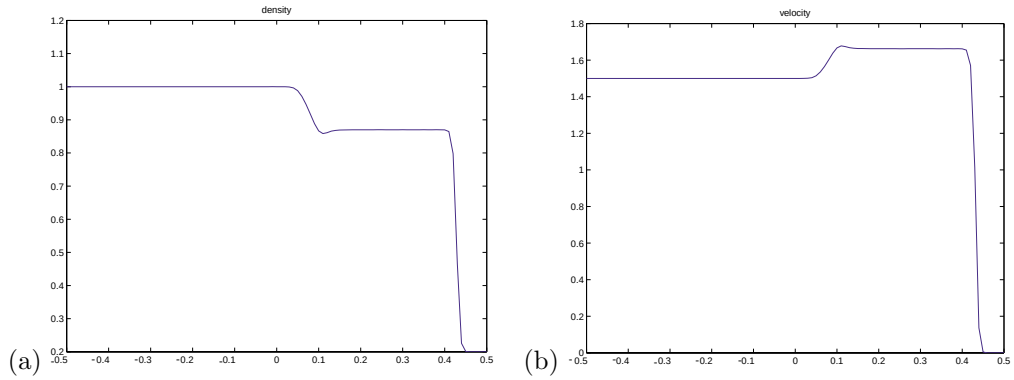


FIG. 5.1. *Density and velocity for* $\epsilon = 1.4$.

5. Formation process of δ-shocks: Numerical simulations. After the cavitation process in the Riemann solutions of (1.1)–(1.3) has been described clearly as the pressure decreases in section 4, understanding the formation process of δ-shocks in the Riemann solutions as the pressure decreases becomes more constructive for comparison. For this purpose, in this section we present a selected group of representative numerical results in the level of Euler dynamics (1.1)–(1.3) starting with Riemann initial data. We have performed many more numerical tests to make sure what we present are not numerical artifacts.

We solve the Riemann problem for (1.1)–(1.3) with $p(\rho) = \rho^\gamma/\gamma$ and $\gamma = 1.4$ for an ideal gas. The Riemann initial data is

$$(\rho, v)(x, 0) = \begin{cases} (1.0, 1.5) & \text{for} \quad x < 0, \\ (0.2, 0.0) & \text{for} \quad x > 0. \end{cases}$$

To discretize the system, we use the higher order ENO scheme to obtain a method-of-line ordinary differential equation in time and then discretize the ordinary differential equation by the classical higher order explicit Runge–Kutta method (see [23, 24]). We calculate by the third-order ENO scheme [24] up to $t = 0.2$ with mesh 100. The numerical simulations with different choices of ϵ are presented in Figures 5.1–5.3. These figures show the formation process of a δ-shock in the two-shock Riemann solutions for the Euler equations (1.1)–(1.3) for isentropic fluids as the pressure decreases. We start with $\epsilon/\gamma = 1.0$ and choose then $\epsilon/\gamma = 0.05$ and finally $\epsilon/\gamma = 0.001$. Figures 5.1a–5.3a show the concentration process of the density yielding a weighted δ-measure in the limit, in which the horizontal axis stands for the space variable x and the vertical axis stands for the density. Figures 5.1b–5.3b show the change of the velocity as ϵ decreases yielding a step function in the limit, in which the horizontal axis stands for the space variable x and the vertical axis stands for the velocity.

We can see clearly from these numerical results that, when ϵ decreases, the locations of the two shocks become closer, and the density of the intermediate state increases dramatically, while the velocity is closer to a step function. In the vanishing pressure limit, the two shocks coincide to form, along with the intermediate state, a δ-shock of the transport equations (1.5)–(1.6), while the velocity is a step function.

We remark that it is delicate to calculate solutions of hyperbolic systems of conservation laws that strict hyperbolicity fails, for which the system of pressureless Euler

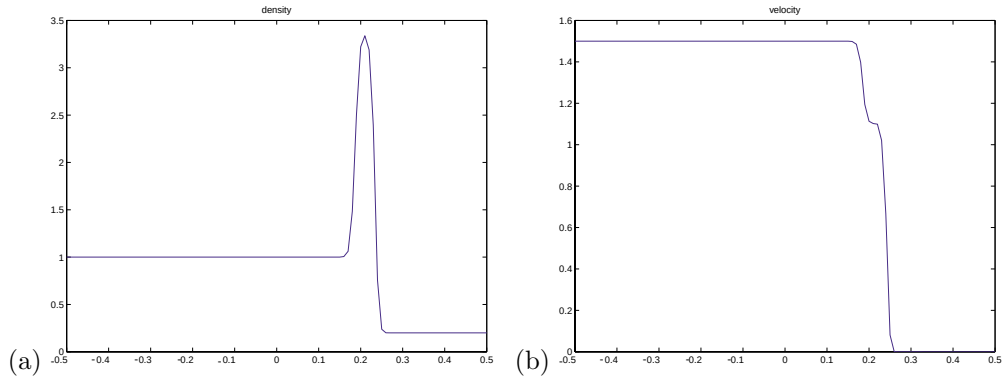


FIG. 5.2. *Density and velocity for* $\epsilon = 0.07$.

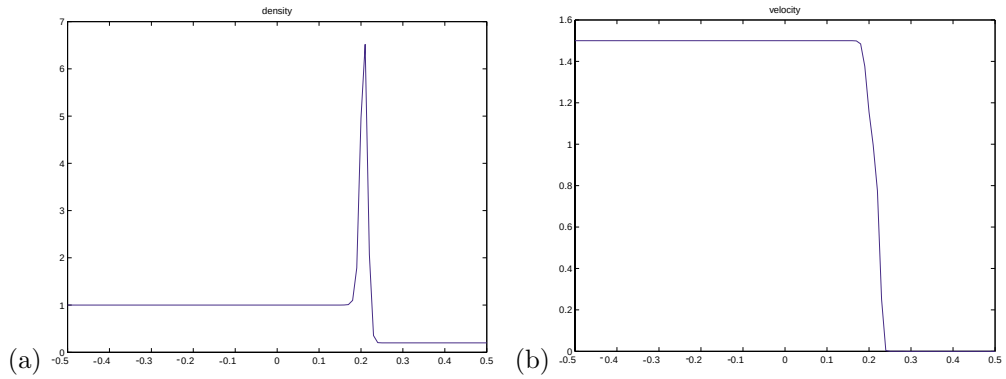


FIG. 5.3. *Density and velocity for* $\epsilon = 0.0014$.

equations for (1.5)–(1.6) is an example. In this section, we have proposed an efficient numerical approach to calculate the solutions containing δ-shocks for (1.5)–(1.6) via the vanishing pressure limit. It would be interesting to apply the approach and ideas set forth here to develop efficient numerical algorithms to calculate solutions for more complex physical models.

REFERENCES

[1] F. BOUCHUT, *On zero pressure gas dynamics*, in Advances in Kinetic Theory and Computing, Ser. Adv. Math. Appl. Sci. 22, World Science Publishing, River Edge, NJ, 1994, pp. 171–190.

[2] F. BOUCHUT AND F. JAMES, *Duality solutions for pressureless gases, monotone scalar conservation laws, and uniqueness*, Comm. Partial Differential Equations, 24 (1999), pp. 2173–2189.

[3] Y. BRENIER AND E. GRENIER, *Sticky particles and scalar conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 2317–2328.

[4] T. CHANG, G.-Q. CHEN, AND S. YANG, 2-*D Riemann problem in gas dynamics and formation of spiral*, in Nonlinear Problems in Engineering and Science—Numerical and Analytical Approach (Beijing, 1991), Science Press, Beijing, 1992, pp. 167–179.

[5] T. CHANG, G.-Q. CHEN, AND S. YANG, *On the Riemann problem for two-dimensional Euler equations* I: *Interaction of shocks and rarefaction waves*, Discrete Contin. Dynam. Systems, 1 (1995), pp. 555–584.

[6] T. Chang, G.-Q. Chen, and S. Yang, *On the Riemann problem for two-dimensional Euler equations* II: *Interaction of contact discontinuities*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 419–430.

[7] G.-Q. Chen and H. Frid, *Divergence-measure fields and hyperbolic conservation laws*, Arch. Ration. Mech. Anal., 147 (1999), pp. 89–118.

[8] G.-Q. Chen and H. Frid, *On the theory of divergence-measure fields and its applications*, Bol. Soc. Brasil. Mat. (N.S.), 32 (2001), pp. 1–33.

[9] G.-Q. Chen and H. Frid, *Extended divergence-measure fields and the Euler equations of gas dynamics*, Comm. Math. Phys., to appear.

[10] G.-Q. Chen and D. Wang, *The Cauchy problem for the Euler equations for compressible fluids*, in Handbook of Mathematical Fluid Dynamics, Vol. 1, Elsevier, Amsterdam, The Netherlands, 2002, pp. 421–543.

[11] W. E, Yu. G. Rykov, and Ya. G. Sinai, *Generalized variational principles, global weak solutions and behavior with random initial data for systems of conservation laws arising in adhesion particle dynamics*, Commun. Math. Phys., 177 (1996), pp. 349–380.

[12] E. Grenier, *Existence globale pour le système des gaz sans pression*, C.R. Acad. Sci. Paris Sér. I. Math., 321 (1995), pp. 171–174.

[13] F. Huang and Z. Wang, *Well posedness for pressureless flow*, Comm. Math. Phys., 222 (2001), pp. 117–146.

[14] K. T. Joseph, *A Riemann problem whose viscosity solutions contain delta measure*, Asymptot. Anal., 7 (1993), pp. 105–120.

[15] B. L. Keyfitz and H. Kranzer, *Spaces of weighted measures for conservation laws with singular shock solutions*, J. Differential Equations, 118 (1995), pp. 420–451.

[16] D. J. Korchinski, *Solutions of a Riemann Problem for a* 2×2 *System of Conservation Laws Possessing No Classical Solution*, thesis, Adelphi University, Garden City, NY, 1977.

[17] J. Li and H. Yang, *Delta-shocks as limits of vanishing viscosity for multidimensional zero-pressure gas dynamics*, Quart. Appl. Math., 59 (2001), pp. 315–342.

[18] J. Li and T. Zhang, *Generalized Rankine-Hugoniot relations of delta-shocks in solutions of transportation equations*, in Advances in Nonlinear Partial Differential Equations and Related Areas (Beijing, 1997), G.-Q. Chen et al., eds., World Science Publishing, River Edge, NJ, 1998, pp. 219–232.

[19] J. Li and T. Zhang, *On the initial-value problem for zero-pressure gas dynamics*, in Hyperbolic Problems: Theory, Numerics, Applications, Vol. 2 (Zürich, 1998), Birkhäuser, Basel, 1999, pp. 629–640.

[20] F. Poupaud and M. Rascle, *Measure solutions to the linear multi-dimensional transport equation with non-smooth coefficients*, Comm. Partial Differential Equations, 22 (1997), pp. 337–358.

[21] M. Sever, *A class of nonlinear, nonhyperbolic systems of conservation laws with well-posed initial value problem*, J. Differential Equations, 180 (2002), pp. 238–271.

[22] W. Sheng and T. Zhang, *The Riemann problem for the transportation equations in gas dynamics*, Mem. Amer. Math. Soc., 137 (1999).

[23] C. W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.

[24] C. W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes* II, J. Comput. Phys., 83 (1989), pp. 32–78.

[25] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, New York, 1994.

[26] D. Tan and T. Zhang, *Two dimensional Riemann problem for a hyperbolic system of nonlinear conservation laws*, J. Differential Equations, 111 (1994), pp. 203–283.

[27] D. Tan, T. Zhang, and Y. Zheng, *Delta-shock waves as limits of vanishing viscosity for hyperbolic systems of conservation laws*, J. Differential Equations, 112 (1994), pp. 1–32.

[28] F. Huang and Z. Wang, *Well posedness for pressureless flow*, Comm. Math. Phys., 222 (2001), pp. 117–146.

[29] Z. Wang, F. Huang, and X. Ding, *On the Cauchy problem of transportation equations*, Acta Math. Appl. Sinica (English Ser.), 13 (1997), pp. 113–122.

[30] Ya. B. Zeldovich, *Gravitational instability: An approximate theory for large density perturbations*, Astronaut. Astrophys., 5 (1970), pp. 84–89.

SIAM J. MATH. ANAL.
Vol. 34, No. 4, pp. 939–956

THE DECREASE OF BULK-SUPERCONDUCTIVITY CLOSE TO THE SECOND CRITICAL FIELD IN THE GINZBURG–LANDAU MODEL*

ETIENNE SANDIER† AND SYLVIA SERFATY‡

Abstract. We study solutions of the Ginzburg–Landau equations describing superconductors in a magnetic field, just below the "second critical field" H_{c_2}. We thus bridge the gap between the situations described in [E. Sandier and S. Serfaty, *Rev. Math. Phys.*, 12 (2000), pp. 1219–1257] and [X. B. Pan, *Comm. Math. Phys.*, 228 (2002), pp. 327–370]. We prove estimates on the energy, among them one by an algebraic trick inspired by the Bogomoln'yi trick for self-duality. We thus show how, for energy-minimizers, superconductivity decreases in average in the bulk of the sample when the applied field increases to H_{c_2}.

Key words. superconductivity, second critical field, phase transitions, asymptotic analysis

AMS subject classifications. 82D55, 35B40, 35B25, 35J60, 35J20, 35Q99, 58E50

PII. S0036141002406084

1. Introduction.

Superconductivity is modeled by the two-dimensional Ginzburg–Landau free energy

$$J(u,A)=\frac{1}{2}\int_\Omega |\nabla_A u|^2+|h-h_{\rm ex}|^2+\frac{\kappa^2}{2}(1-|u|^2)^2. \tag{1.1}$$

We are interested in studying critical points of this energy when the applied field $h_{\rm ex}$ gets close (from below) to the "second critical field" H_{c_2}.

Let us first explain the notation. Ω is a smooth, bounded, simply connected domain of $\mathbb{R}^2$, corresponding to the section of an infinite cylindrical body. J is a function of u, the "order parameter," complex-valued function, and of the "vector potential" $A:\Omega\mapsto\mathbb{R}^2$. u indicates the local state of the material, $|u|^2\le 1$ being the local density of superconducting electrons. Roughly speaking, where $|u|\sim 1$ it is the "superconducting phase," while where $|u|\sim 0$ it is the "normal phase." A is the potential associated to the magnetic field $h=\operatorname{curl} A=\partial_1 A_2-\partial_2 A_1$ (real-valued function) that exists in the sample. ∇_A denotes the covariant derivation $\nabla-iA$; it is an abelian gauge theory, and everything is invariant under gauge transformations: $u\to ue^{i\Phi}, A\to A+\nabla\Phi$. A configuration is really an orbit of gauge-equivalent couples (u,A).

The parameter $h_{\rm ex}$ is the intensity of the applied magnetic field (assumed to be uniform, parallel to the cylinder axis). Finally κ is the Ginzburg–Landau parameter; it is the ratio of two characteristic lengths of the material.

*Received by the editors April 19, 2002; accepted for publication (in revised form) October 31, 2002; published electronically April 9, 2003.
http://www.siam.org/journals/sima/34-4/40608.html
†Département de Mathématiques, Université Paris-12 Val-de-Marne, 61 ave du Général de Gaulle, 94010 Créteil Cedex, France (sandier@univ-paris12.fr).
‡Courant Institute of Mathematical Sciences, 251 Mercer St., New York, NY 10012 (serfaty@cims.nyu.edu). This author was supported by the CNRS.

The equations associated to this functional are the Ginzburg–Landau equations

$$-\nabla_A^2 u = \kappa^2 u(1-|u|^2) \text{ in } \Omega, \tag{1.2}$$

$$-\nabla^\perp h = \langle iu, \nabla_A u\rangle \text{ in } \Omega, \tag{1.3}$$

$$h = h_{\text{ex}} \text{ on } \partial\Omega, \tag{1.4}$$

$$(\nabla u - iAu)\cdot \nu = 0 \text{ on } \partial\Omega, \tag{1.5}$$

where $\nabla^\perp$ denotes $(-\partial_2, \partial_1)$ and $\langle ., .\rangle$ is the scalar product in $\mathbb{C}$ identified with $\mathbb{R}^2$.

When type-II superconductors are submitted to a magnetic field, they exhibit phase transitions for certain critical fields, denoted H_{c_1}, H_{c_2}, and H_{c_3}. When $h_{\text{ex}} \leq H_{c_1}$, the sample is in the superconducting phase everywhere and repels the magnetic field (called the Meissner effect). At H_{c_1}, there is a phase transition where vortices appear. Vortices are zeros of the order parameter u around which u has a nonzero winding number. (For a mathematical description of vortices in Ginzburg–Landau without magnetic field, see [BBH] and subsequent works.) As h_{ex} increases, vortices get more and more numerous and tend to arrange in a triangular lattice, called an "Abrikosov lattice." When $H_{c_2} \leq h_{\text{ex}} \leq H_{c_3}$, the material is in the normal phase everywhere except on a layer near the boundary where superconductivity persists, while for $h_{\text{ex}} \geq H_{c_3}$ it is normal everywhere ($u \equiv 0$). For a more thorough physical presentation, one may see [DeG, SST, T].

We are interested in the "London limit" $\kappa \to +\infty$. We will also write $\varepsilon = \frac{1}{\kappa}$. ε is the lengthscale of a vortex. Letting $\varepsilon \to 0$ corresponds to having vortices that are small compared to the scale of the sample.

Mathematically, a lot of results have been proved on this functional. Let us start with the situation around the third critical field H_{c_3}. First, observe that there is always a trivial normal solution ($u \equiv 0, h \equiv h_{\text{ex}}$) and that its energy is $\frac{1}{4}|\Omega|\kappa^2$. When the applied field h_{ex} is decreased to H_{c_3}, there is a bifurcation from that normal solution to a branch of solutions with superconductivity on the boundary. This superconductivity actually first appears at H_{c_3} near the point of maximal curvature of the boundary.

The story goes back to Saint-James and de Gennes [SdG] and later Chapman [C], who studied the bifurcation in the half-space based on formal analysis. Rigorously, it was proved by Giorgi and Phillips [GP] that $H_{c_3} = O(\kappa^2)$ and that above H_{c_3} the only solution is the normal one. Then, in the particular case of a disc-domain, Bauman, Phillips, and Tang [BPT] considered radially symmetric solutions bifurcating from eigenfunctions. In a general domain, a formula relating H_{c_3} to the curvature of the boundary, as well as a result showing that eigenfunctions concentrate around the points of maximal curvature of the boundary, was first given by Bernoff and Sternberg in [BS] through a formal analysis, by Del Pino, Felmer, and Sternberg [DFS], and simultaneously by Lu and Pan in [LP3], based on the linear analysis of [LP1, LP2]. Finally, Helffer and Pan, using the analysis of [HM], obtained in [HP] the most accurate result, stating that superconductivity first appears at H_{c_3} near the point of maximum curvature of the boundary and that

$$H_{c_3} \sim_{\kappa\to\infty} \frac{\kappa^2}{\beta_0} + \left(\frac{C_1}{\beta_0^{3/2}} k_{max}\right)\kappa, \tag{1.6}$$

where β_0 is the lowest eigenvalue of a Schrödinger operator with magnetic field in the half-plane, and k_{max} is the maximum of the curvature on the boundary of Ω.

Let us now turn to the situation further below H_{c_3}. Recently, Pan proved in [P1] a very nice result describing global minimizers of the energy between H_{c_2} and H_{c_3}. He showed that H_{c_2} can be defined as the infimum of h_{ex} such that global minimizers of J do not have *bulk-superconductivity* but only surface-superconductivity, and that

$$H_{c_2} \sim_{\kappa\to\infty} \kappa^2. \tag{1.7}$$

Following his notation, we define b by

$$h_{\text{ex}} = (b + o(1))\kappa^2, \tag{1.8}$$

and we will also denote by J_D the Ginzburg–Landau functional restricted to a subdomain D of Ω.

He proved the following.

THEOREM 1.1 (Pan [P1]). *Let (u, A) be a minimizer of J. For $1 < b < \frac{1}{\beta_0}$, there exist positive numbers E_b and κ_b such that for $\kappa > \kappa_b$,*

$$J(u, A) \sim_{\kappa\to\infty} \frac{|\Omega|}{4}\kappa^2 - \kappa E_b|\partial\Omega| + o(\kappa), \tag{1.9}$$

where $|\Omega|$ denotes the volume of Ω and $|\partial\Omega|$ denotes the length of $\partial\Omega$. For any closed subdomain D of $\overline{\Omega}$, for $\kappa > \kappa_D$,

$$J_D(u, A) \sim_{\kappa\to\infty} \frac{|D \cap \Omega|}{4}\kappa^2 - \kappa E_b|D \cap \partial\Omega| + o(\kappa). \tag{1.10}$$

Moreover, $\frac{1}{\kappa}|\nabla_A u|$ and $|u|$ exponentially decay in the interior of Ω in the sense that for all $\alpha > 0$, for $\kappa > \kappa(\alpha)$,

$$\int_\Omega \left(|u|^2 + \frac{1}{\kappa^2}|\nabla_A u|^2\right) \exp(\alpha\kappa dist(x, \partial\Omega))\, dx \le \frac{O(1)}{\kappa}.$$

He also proved results for the case $b = 1$. Let us point out that slightly stronger exponential-decay results have been proved by Almog in [Al1] replacing the large-kappa limit by the large-domain limit.

Thus, when h_{ex} is decreased and crosses H_{c_3}, superconductivity first nucleates at the points of maximal curvature of the boundary, and u is a small perturbation of the normal solution 0. As h_{ex} further decreases, a uniform superconducting sheath of scale $\varepsilon = \frac{1}{\kappa}$ rapidly forms on the entire boundary of the sample, while the bulk remains normal as shown in the previous theorem. Superconductivity increases on the boundary as $b \to 1$.

On the other hand, the situation is also well understood for small applied fields: the superconducting state below H_{c_1} has been studied in [S1, S3, SS1]; the value of H_{c_1} being asymptotic to $C(\Omega)\log\kappa$ was proved in [S1, SS1, SS5]. Above H_{c_1}, we showed vortices appear first near the center of the domain [S1, S2], and a vortex region where the density of vortices is uniform and proportional to h_{ex}, surrounded by a purely superconducting region, forms and inflates (see [SS3]). As soon as $h_{\text{ex}} \gg \log\kappa$, the vortex region covers up the whole sample, and we have proved the following.

THEOREM 1.2 (Sandier and Serfaty [SS2]). *Assume h_{ex} is any function of κ such that $\log\kappa \ll h_{\text{ex}} \ll \kappa^2$ as $\kappa \to \infty$. If (u, A) is a corresponding minimizer of J, then*

$$J(u, A) \sim_{\kappa\to\infty} \frac{1}{2}|\Omega| h_{\text{ex}} \log\frac{\kappa}{\sqrt{h_{\text{ex}}}}, \tag{1.11}$$

where $|\Omega|$ denotes the volume of Ω; and if D is any closed subdomain of $\overline{\Omega}$, then

$$J_D(u,A) \sim_{\kappa\to\infty} \frac{1}{2}|D|h_{\rm ex}\log\frac{\kappa}{\sqrt{h_{\rm ex}}}. \tag{1.12}$$

Moreover, the density of vortices converges in some sense to the uniform density $h_{\rm ex}$.

In this regime $h_{\rm ex}\ll\kappa^2$, the superconducting phase surrounding the vortices still dominates in the sense that, from estimate (1.11), $\int_\Omega(1-|u|^2)^2=o(1)$. Essentially, one can think of the vortices as of degree 1 and placed regularly, for example, on a periodic lattice, one per cell of size $\frac{1}{\sqrt{h_{\rm ex}}}$, which remains much larger than their characteristic size ε (as long as $h_{\rm ex}\ll\kappa^2$).

The question is thus to bridge the gap between the situations of these two theorems (that of $h_{\rm ex}\ll\kappa^2$, i.e., $b=0$, and that above H_{c_2}, i.e., $b>1$) in the only range of applied fields which remained unstudied: $b\in[0,1]$. How do the vortices disappear and how does the bulk superconductivity disappear? Essentially two scenarios could be suggested. One is that as $h_{\rm ex}$ increases, the distance between the vortices decreases, and before it becomes smaller than their size $O(\varepsilon)$, the vortices merge into one "giant vortex" of large degree. The other scenario is that $\max|u|$ decreases in the bulk, while the vortex array structure remains unchanged, until $|u|$ is close to 0 in the bulk, and superconductivity remains only on the boundary, as described by Pan. It is considered by physicists that it is the second scenario rather than the first which occurs, at least in this limit $\kappa\to\infty$, and the results we prove confirm this. However, giant vortices do occur (and are observed) for smaller κ.

We start with a general lower bound result, proved through a very simple argument. Introducing the operator $\mathcal{D}_A=\partial_1+i\partial_2-i(A_1+iA_2)$, we have the identity

$$|\mathcal{D}_A u|^2=|\nabla_A u|^2-\operatorname{curl}(iu,\nabla_A u)-|u|^2h. \tag{1.13}$$

This operator is the one that was used by Bogomoln'yi (see [JT]) to exhibit the self-duality of the Ginzburg–Landau equations for $\kappa=\frac{1}{\sqrt{2}}$. By a purely algebraic manipulation quite similar to the trick of Bogomoln'yi (the same kind of manipulation was also behind the results of [M] and [GP]), we deduce from (1.13) the following *nontrivial* lower bound.

PROPOSITION 1.3. *Let (u,A) be any solution of the Ginzburg–Landau system (1.2)–(1.5) and $B\subset\Omega$ any ball of radius $R\gg\varepsilon=\frac{1}{\kappa}$. Then if $b\geq 1$,*

$$\frac{J_B(u,A)}{|B|}\geq\frac{\kappa^2}{4}+o(\kappa^2), \tag{1.14a}$$

while if $b\leq 1$,

$$\begin{aligned}\frac{J_B(u,A)}{|B|}&=\left(\frac{b}{2}-\frac{b^2}{4}\right)\kappa^2+\frac{1}{2|B|}\int_B|\mathcal{D}_A u|^2+\frac{\kappa^2}{2}(1-b-|u|^2)^2+|h-h_{\rm ex}|^2+o(\kappa^2)\\&\geq\left(\frac{b}{2}-\frac{b^2}{4}\right)\kappa^2+o(\kappa^2).\end{aligned} \tag{1.14b}$$

By this we mean that given a function $R(\kappa)\gg 1/\kappa$, there exists a function $o(\kappa^2)$ such that the above inequalities are verified for any solution of (1.2)–(1.5).

Observe that this estimate is true for *any solution* of the equations, not necessarily minimizing or stable. It is in fact true for any configuration that satisfies the a priori estimates $\|u\|_{L^\infty(\overline{\Omega})}\leq 1$, $\|\nabla_A u\|_{L^\infty(\overline{\Omega})}\leq C\kappa$, $\|h-h_{\rm ex}\|_{L^\infty(\overline{\Omega})}\leq C\kappa$ (see the proof).

Let us now turn to energy-minimizers. We denote by $\min J_{B_R}$ the minimum of the energy-functional on a ball B_R, i.e.,

$$\min J_{B_R} = \min_{(u,A)} \frac{1}{2} \int_{B_R} |\nabla_A u|^2 + |\operatorname{curl} A - h_{\text{ex}}|^2 + \frac{\kappa^2}{2}(1-|u|^2)^2.$$

We also denote by $(\overline{u}, \overline{A})$ a minimizer for this problem.

THEOREM 1.4. *Let $0 \leq b \leq 1$. There exists a continuous increasing function f from $[0,1]$ to $[0, \frac{1}{4}]$ such that, as $\kappa \to \infty$, for (u,A) any minimizer of J, for all $R_\kappa \gg \varepsilon = \frac{1}{\kappa}$, and for all balls B_{R_κ} in Ω,*

$$\frac{J_{B_R}(u,A)}{\kappa^2 |B_R|} \sim \frac{\min J_{B_R}}{\kappa^2 |B_R|} \longrightarrow f(b), \tag{1.15}$$

$$\frac{1}{|B_R|} \int_{B_R} |u|^4 \sim \frac{1}{|B_R|} \int_{B_R} |\overline{u}|^4 \longrightarrow 1 - 4f(b), \tag{1.16}$$

$$|u|^4 \rightharpoonup 1 - 4f(b) \quad \text{in } L^\infty \text{ weak-}*, \tag{1.17}$$

$$\frac{1-4f(b)}{1-b} - o(1) \leq \frac{1}{|B_R|} \int_{B_R} |u|^2 \leq \sqrt{1-4f(b)} + o(1), \tag{1.18}$$

and the following estimates hold: There exists universal constants $0 < \alpha < 1$ and $c > 0$ such that

$$\frac{b}{2} - \frac{b^2}{4} \leq f(b) \leq \min\left(\frac{b}{4}\left(\log \frac{1}{b} + c \right), \frac{1-\alpha(1-b)^2}{4} \right) \leq \frac{1}{4}, \tag{1.19}$$

and hence

$$\alpha(1-b)^2 \leq 1 - 4f(b) \leq (1-b)^2.$$

COROLLARY 1.5. *For all D closed subdomains of $\overline{\Omega}$,*

$$J_D(u,A) \sim_{\kappa \to \infty} |D| f(b) \kappa^2.$$

COROLLARY 1.6. *$f(0) = 0$ and $f(1) = \frac{1}{4}$. Therefore for $b = 0$ and for all $R_\kappa \gg \varepsilon$ we get the following result proved in* [SS2]:

$$\lim_{\kappa \to \infty} \frac{J_{B_R}(u,A)}{\kappa^2 |B_R|} = 0.$$

For $b \geq 1$ and for all $R_\kappa \gg \varepsilon$ we get

$$\lim_{\kappa \to \infty} \frac{J_{B_R}(u,A)}{\kappa^2 |B_R|} = \frac{1}{4},$$

which follows also from [P1] *and Proposition* 1 *in* [SS2].

We thus show that the loss of superconductivity happens through a decrease of the average of $|u|^4$, such as $(1-b)^2$ in Ω, and that the energy-repartition remains uniform. Those two facts go in the direction of the second scenario.

We have given asymptotic estimates of the minimal energy which extend that of (1.11). We have proved that the energy is uniformly spread over Ω and that a minimizer almost minimizes locally the energy at any scale $\gg \varepsilon$ (ε being the characteristic

scale of variation of u). At scales $O(\varepsilon)$ this ceases to be true: minimizers in regions of smaller sizes start to depend greatly on the region-size, as seen in [AD]. We have also shown that for global minimizers, some superconductivity remains in the bulk as long as $b < 1$, since from (1.18) the average of $|u|^2$ remains larger than $\alpha(1-b)$.

This theorem relies on upper and lower bounds for the energy, in the spirit of gamma-convergence. It seems difficult to give a more explicit expression or a finer estimate on $f(b)$. The lower bound is given by Proposition 1.3. The upper bound is obtained by constructing test configurations. They are chosen to be periodic with respect to a square lattice of size $\sqrt{\frac{2\pi}{b}}\varepsilon$, with a vortex of degree 1 in each cell. In view of (1.14), a minimizer (u, A) should almost be a minimizer of $\frac{1}{2}\int_\Omega |\mathcal{D}_A u|^2 + |h - b\kappa^2|^2 + \frac{\kappa^2}{2}(1-b-|u|^2)^2$. We choose our test configuration to satisfy $|u| \leq C\sqrt{1-b}$ and also $\mathcal{D}_A u = 0$ (following somewhat the construction of [JT] of vortex solutions in the self-dual case). This configuration of course has no reason to be optimal (nor does the square-shape) but gives the right order of energy.

We get as a corollary of Proposition 1.3 and Theorem 1.4 that, for energy-minimizers,

$$\begin{aligned}\limsup_{\kappa\to\infty} \frac{1}{2\kappa^2|B_R|}\int_{B_R} |\mathcal{D}_A u|^2 + \frac{\kappa^2}{2}(1-b-|u|^2)^2 + |h-h_{\text{ex}}|^2 \leq\ & f(b) - \frac{b}{2} + \frac{b^2}{4} \\ \leq\ & \frac{1-\alpha}{4}(1-b)^2\end{aligned}$$

and that

$$\frac{1}{R^2}\int_{B_R} (1-b-|u|^2)^2 \leq (1-\alpha)(1-b)^2, \tag{1.20}$$

from which one can deduce (1.18). It is also tempting, in view of (1.14), to think that $|u|^2 \leq C(1-b)$ in the bulk.

There remain many open questions on the behavior of minimizers, which all seem quite delicate.

First of all, we conjecture that, next to an interior point of Ω, a minimizing solution should converge, after blow-up at the scale ε, to a unique limiting profile in $\mathbb{R}^2$. A much more difficult task would be to show that this limiting profile is periodic. For a study of periodic solutions of Ginzburg–Landau, see [Du], [C], and [Al2].

We have not mentioned vortices of the minimizers. It is difficult to describe them and even define them: $|u|$ becomes uniformly small, so one can no longer define the vortices as the regions where $|u|$ is small. Nevertheless, there should be vortices (they appear in our upper bound construction), with a total degree $2\pi h_{\text{ex}}$ on the boundary of Ω. Heuristically, using the second Ginzburg–Landau equation,

$$-\frac{\nabla^\perp h}{\rho^2} + h = \nabla\varphi,$$

where we write $u = \rho e^{i\varphi}$ in polar coordinates. Taking the curl of this equation, we are led to

$$\text{div}\ \left(\frac{\nabla h}{\rho^2}\right) + h = \pi\sum_i d_i\delta_{a_i},$$

where the a_i are the zeros (or vortices) of u, and d_i their degrees or winding number.

Since $h \to h_{\rm ex}$ strongly, we should have, at least formally,

$$2\pi \sum_i d_i \delta_{a_i} \sim h_{\rm ex}$$

(as we had for $h_{\rm ex} \ll \kappa^2$). However, it seems difficult to give a rigorous meaning to this statement. We can prove that on any subdomain D of volume $R^2 \gg \frac{1}{\kappa^2}$ such that $|u| > c > 0$ independently of κ on ∂D, and such that the perimeter of D is less than $O(R)$, the total degree of u on ∂D is equivalent to $h_{\rm ex}|D|$. But the existence of such a D is not proved.

To conclude, it would be very nice, but certainly difficult, to prove a bifurcation at H_{c_2} from the surface-superconductivity solution to one of the known periodic-like vortex solutions.

2. The algebraic trick. From now on, we denote $h = \operatorname{curl} A$ and $u = \rho e^{i\varphi}$ in polar coordinates. Then

$$|\nabla_A u|^2 = |\nabla \rho|^2 + \rho^2 |\nabla \varphi - A|^2.$$

We are interested in this section in studying families of solutions of Ginzburg–Landau, or configurations which satisfy the following a priori estimates.

LEMMA 2.1. *If (u, A) is a solution of Ginzburg–Landau, we have*

$$\|h - h_{\rm ex}\|_{C^1(\overline{\Omega})} \le C\kappa, \qquad \|h - h_{\rm ex}\|_{C^2(\overline{\Omega})} \le C\kappa^2, \tag{2.1}$$

$$\|\nabla_A u\|_{L^\infty(\overline{\Omega})} \le C\kappa, \qquad \|\nabla \rho\|_{L^\infty(\overline{\Omega})} \le C\kappa, \tag{2.2}$$

$$e_\kappa(u, A) := |\nabla_A u|^2 + |h - h_{\rm ex}|^2 + \frac{\kappa^2}{2}(1 - \rho^2)^2 \le C\kappa^2. \tag{2.3}$$

These estimates are proved in [HP, Proposition 4.3]; see also [P1, Lemma 7.1]. They rely on a blow-up at scale $\varepsilon = \frac{1}{\kappa}$, which leads to equations at scale 1, for which all the quantities are uniformly bounded.

Proof of Proposition 1.3. As already mentioned, the proof relies on the Bogomoln'yi identity on the operator $\mathcal{D}_A = \partial_1 + i\partial_2 - i(A_1 + iA_2)$. One can check that, in polar coordinates,

$$|\mathcal{D}_A u|^2 = |\rho(\nabla \varphi - A) - \nabla^\perp \rho|^2. \tag{2.4}$$

Expanding the square on the right-hand side, one gets the crucial identity

$$|\mathcal{D}_A u|^2 = |\nabla_A u|^2 - \operatorname{curl} j - \rho^2 h, \tag{2.5}$$

where j is the superconducting current $\langle iu, \nabla_A u \rangle$. Inserting (2.5) in J, we are led to

$$J_{B_R}(u, A) = \frac{1}{2} \int_{B_R} |\mathcal{D}_A u|^2 + \operatorname{curl} j + \rho^2 h + |h - h_{\rm ex}|^2 + \frac{\kappa^2}{2}(1 - \rho^2)^2. \tag{2.6}$$

Moreover, using the fact that $|j| \le |\nabla_A u| \le C\kappa$ with (2.2), we have

$$\left| \int_{B_R} \operatorname{curl} j \right| = \left| \int_{\partial B_R} j \cdot \tau \right| \le \int_{\partial B_R} |j| \le O(R\kappa). \tag{2.7}$$

Also $h = h_{\rm ex} + O(\kappa) = b\kappa^2 + o(\kappa^2)$ in view of (2.1). Combining these facts with (2.6) yields

(2.8)
$$\begin{aligned} J_{B_R}(u,A) &= \frac{1}{2}\int_{B_R} |\mathcal{D}_A u|^2 + \rho^2 b\kappa^2 + \frac{\kappa^2}{2}(1-\rho^2)^2 + |h-h_{\rm ex}|^2 + O(R\kappa) + o(R^2\kappa^2) \\ &= \frac{1}{2}\int_{B_R} |\mathcal{D}_A u|^2 + \kappa^2\left(\frac{1}{2} + \rho^2(b-1) + \frac{\rho^4}{2}\right) + |h-h_{\rm ex}|^2 + O(R\kappa) + o(R^2\kappa^2). \end{aligned}$$

If $b \geq 1$, this immediately implies that

$$\frac{J_{B_R}(u,A)}{\kappa^2|B_R|} \geq \frac{1}{4} + o(1).$$

(Thus, we see why the value $b = 1$ plays a particular role.)

If $b \leq 1$, we observe that $\rho^2(b-1) + \frac{\rho^4}{2} = \frac{1}{2}(\rho^2 - (1-b))^2 - \frac{1}{2}(1-b)^2$ and obtain

(2.9)
$$\begin{aligned} J_{B_R}(u,A) = |B_R|\frac{\kappa^2}{4}(1-(1-b)^2) + \frac{1}{2}\int_{B_R} |\mathcal{D}_A u|^2 + \frac{\kappa^2}{2}(1-b-\rho^2)^2 + |h-h_{\rm ex}|^2 \\ + O(R\kappa) + o(R^2\kappa^2). \end{aligned}$$

We conclude that (1.14) holds. ∎

3. Energy localization and convergence. We are now interested in families of global minimizers of J. The following lemma allows us to localize all energy comparisons.

LEMMA 3.1. *Let R_κ be such that $R_\kappa \gg \varepsilon$. Then, (u,A) being a minimizer of J, for any ball B_R of radius R_κ in Ω,*

$$\frac{J_{B_R}(u,A)}{\kappa^2|B_R|} = \frac{\min J_{B_R}}{\kappa^2|B_R|} + o(1).$$

Proof. One inequality is obvious:

$$J_{B_R}(u,A) \geq \min J_{B_R}.$$

The converse relies on a comparison argument. Let $(\tilde{u},\tilde{A})$ be a minimizer of J_{B_R}. We construct a test configuration in Ω which coincides with (u,A) in $\Omega\backslash B_R$, and with $(\tilde{u},\tilde{A})$ in $B_{R-3\varepsilon}$.

Let χ be a $C^\infty(\Omega)$ function such that

(3.1)
$$\begin{cases} \chi(x) = 1 & \text{in } \Omega\backslash B_R, \\ \chi(x) = 0 & \text{in } B_{R-\frac{\varepsilon}{2}}\backslash B_{R-\frac{5}{2}\varepsilon}, \\ \chi(x) = 1 & \text{in } B_{R-3\varepsilon}, \\ |\nabla\chi| \leq \dfrac{C}{\varepsilon}, \\ \displaystyle\int_\Omega |\nabla\chi|^2 \leq O\left(\frac{R}{\varepsilon}\right). \end{cases}$$

We define $(\overline{u}, \overline{A})$ by

$$(\overline{u}, \overline{A}) = (\chi u, A) \text{ in } \Omega \backslash B_{R-\varepsilon},$$
$$(\overline{u}, \overline{A}) = (\chi \tilde{u}, \tilde{A}) \text{ in } B_{R-2\varepsilon}.$$

There remains the matter of extending $(\overline{u}, \overline{A})$ in $B_{R-\varepsilon} \backslash B_{R-2\varepsilon}$. We take $\overline{u} = 0$ there and may extend $\overline{A}$ in such a way that

$$\|\operatorname{curl} \overline{A} - h_{\text{ex}}\|_{L^\infty(\Omega)} \leq C\kappa \tag{3.2}$$

(indeed, this is true for A and $\tilde{A}$.) Then, (u, A) being a minimizer of J, we have

$$\begin{aligned} 0 \geq J(u, A) - J(\overline{u}, \overline{A}) &= \int_{B_R} e_\kappa(u, A) - e_\kappa(\overline{u}, \overline{A}) \\ &= \int_{B_R \backslash B_{R-\varepsilon}} + \int_{B_{R-\varepsilon} \backslash B_{R-2\varepsilon}} + \int_{B_{R-2\varepsilon} \backslash B_{R-3\varepsilon}} + \int_{B_{R-3\varepsilon}} e_\kappa(u, A) - e_\kappa(\overline{u}, \overline{A}). \end{aligned} \tag{3.3}$$

Then

$$\begin{aligned} &\left| \int_{B_R \backslash B_{R-\varepsilon}} e_\kappa(u, A) - e_\kappa(\overline{u}, \overline{A}) \right| \\ &\qquad = \frac{1}{2} \left| \int_{B_R \backslash B_{R-\varepsilon}} |\nabla \rho|^2 + \rho^2 |\nabla \varphi - A|^2 + \frac{\kappa^2}{2}(1 - \rho^2)^2 \right. \\ &\qquad\qquad \left. - \int_{B_R \backslash B_{R-\varepsilon}} |\nabla(\chi\rho)|^2 + \chi^2 \rho^2 |\nabla \varphi - A|^2 + \frac{\kappa^2}{2}(1 - \rho^2 \chi^2)^2 \right| \\ &= \frac{1}{2} \left| \int_{B_R \backslash B_{R-\varepsilon}} (1 - \chi^2) |\nabla_A u|^2 + \frac{\kappa^2}{2} \left((1 - \rho^2)^2 - (1 - \rho^2 \chi^2)^2 \right) - \int_{B_R \backslash B_{R-\varepsilon}} \rho^2 |\nabla \chi|^2 \right| \\ &\leq O\left(\frac{R}{\varepsilon} \right), \end{aligned} \tag{3.4}$$

where we have used (2.3) and (3.1). Similarly, exchanging the roles of (u, A) and $(\tilde{u}, \tilde{A})$, we find

$$\left| \int_{B_{R-2\varepsilon} \backslash B_{R-3\varepsilon}} e_\kappa(u, A) - e_\kappa(\overline{u}, \overline{A}) \right| \leq O\left(\frac{R}{\varepsilon} \right). \tag{3.5}$$

In $B_{R-\varepsilon} \backslash B_{R-2\varepsilon}$, $\overline{u} = 0$, so with (2.3) again and (3.2),

$$\left| \int_{B_{R-\varepsilon} \backslash B_{R-2\varepsilon}} e_\kappa(u, A) - e_\kappa(\overline{u}, \overline{A}) \right| \leq O\left(\frac{R}{\varepsilon} \right) + \frac{1}{2} \int_{B_{R-\varepsilon} \backslash B_{R-2\varepsilon}} |\operatorname{curl} \overline{A} - h_{\text{ex}}|^2 \leq O\left(\frac{R}{\varepsilon} \right). \tag{3.6}$$

By (2.3) again, we have

$$\int_{B_{R-3\varepsilon}} e_\kappa(u, A) = \int_{B_R} e_\kappa(u, A) + O\left(\frac{R}{\varepsilon} \right), \tag{3.7}$$

$$\int_{B_{R-3\varepsilon}} e_\kappa(\tilde{u}, \tilde{A}) = \int_{B_R} e_\kappa(\tilde{u}, \tilde{A}) + O\left(\frac{R}{\varepsilon} \right). \tag{3.8}$$

But, since $(\tilde{u}, \tilde{A})$ minimizes J_{B_R}, we have

$$\int_{B_R} e_\kappa(u, A) \geq \int_{B_R} e_\kappa(\tilde{u}, \tilde{A}). \tag{3.9}$$

Combining this with (3.7) and (3.8), and using the fact that $(\overline{u}, \overline{A})$ is equal to $(\tilde{u}, \tilde{A})$ in $B_{R-3\varepsilon}$, we deduce that

$$\int_{B_{R-3\varepsilon}} e_\kappa(u, A) - e_\kappa(\overline{u}, \overline{A}) \geq O\left(\frac{R}{\varepsilon}\right).$$

Combining this with (3.3)–(3.7), we get

$$\left| \int_{B_R} e_\kappa(u, A) - e_\kappa(\overline{u}, \overline{A}) \right| \leq O\left(\frac{R}{\varepsilon}\right),$$

i.e.,

$$J_{B_R}(u, A) = J_{B_R}(\tilde{u}, \tilde{A}) + O(R\kappa),$$

which leads to the result. □

What we did with balls in the previous lemma can be done with squares K_R of size R.

LEMMA 3.2. *For all $b \geq 0$, and for $R_\kappa \geq R'_\kappa \gg \varepsilon$,*

$$\frac{\min J_{K_R}}{\kappa^2 |K_R|} = \frac{\min J_{K_{R'}}}{\kappa^2 |K_{R'}|} + o(1); \tag{3.10}$$

hence $\frac{\min J_{K_R}}{\kappa^2 |K_R|}$ does not depend on $R \gg \varepsilon$ (up to an $o(1)$).

Proof. Let us denote by [.] the integer part of a real number. Assume first that $R' \ll R$. K_R can be split into at least $[R^2/R'^2]$ disjoint squares of size R'. Thus, for (u, A) a minimizer of J_{K_R},

$$\begin{aligned} J_{K_R}(u, A) &\geq \left[\frac{R^2}{(R')^2}\right] J_{K_{R'}}(u, A) \\ &\geq \left[\frac{R^2}{(R')^2}\right] \min J_{K'_R}. \end{aligned}$$

We deduce that

$$\frac{\min J_{K_R}}{\kappa^2 |K_R|} \geq \frac{\min J_{K_{R'}}}{\kappa^2 |K_{R'}|}(1 + o(1)).$$

Conversely, let us split K_R into $[R^2/(R')^2] + o(1)$ squares of size R' with a layer of size 3ε between them. Using the pasting procedure of Lemma 3.1, we can construct a test configuration (u, A) in K_R that agrees with the minimizer of $J_{K_{R'}}$ in each subsquare of size R', and such that

$$J_{K_R}(u, A) \leq ([R^2/(R')^2] + o(1)) \left(\min J_{K_{R'}} + C\frac{R'}{\varepsilon}\right).$$

We can check that the error terms are negligible and deduce that

$$\frac{\min J_{K_R}}{\kappa^2 |K_R|} \leq \frac{\min J_{K_{R'}}}{\kappa^2 |K_{R'}|}(1 + o(1)).$$

Since for all R, $\frac{\min J_{K_R}}{\kappa^2|K_R|} \leq \frac{1}{4} \leq O(1)$ (by comparison with the normal solution), we deduce that (3.10) holds. If R and R' are of the same order, one may introduce R'' such that $R' \gg R'' \gg \varepsilon$. From the above, one deduces that

$$\frac{\min J_{K'_R}}{\kappa^2|K'_R|} = \frac{\min J_{K_{R''}}}{\kappa^2|K_{R''}|} + o(1),$$

and the same with R' replaced by R, from which it follows that

$$\frac{\min J_{K_R}}{\kappa^2|K_R|} = \frac{\min J_{K'_R}}{\kappa^2|K'_R|} + o(1). \qquad \square$$

LEMMA 3.3. *For all $R_\kappa \gg \varepsilon$, $\frac{\min J_{B_R}}{\kappa^2|B_R|}$ has a limit as $\kappa \to \infty$, which depends only on b. We denote it by $f(b)$. f is continuous, increasing in $[0,1]$.*

Proof. Consider $R_\kappa \gg \varepsilon$ and (u, A) as a minimizer of J_{B_R}. We denote for a moment by J_{κ,B_R} the functional for κ defined on B_R (b being fixed, $h_{\text{ex}} = b\kappa^2$). Let $\lambda < 1$, and define in $B_{\frac{R}{\lambda}}$,

$$v(x) = u(\lambda x), \qquad B(x) = \lambda A(\lambda x).$$

Then, by change of variables, we have

(3.11)
$$\min J_{\kappa,B_R} = J_{\kappa,B_R}(u,A) = \frac{1}{2}\int_{B_{R/\lambda}} |\nabla_B v|^2 + \frac{1}{\lambda^2}|\operatorname{curl} B - \kappa^2\lambda^2 b|^2 + \frac{\kappa^2\lambda^2}{2}(1-|v|^2)^2.$$

Since $\frac{1}{\lambda} > 1$, this implies that

$$\begin{aligned}
\frac{\min J_{\kappa,B_R}}{\kappa^2R^2} &\geq \frac{1}{2\kappa^2R^2}\int_{B_{R/\lambda}} |\nabla_B v|^2 + \frac{1}{\lambda^2}|\operatorname{curl} B - \kappa^2\lambda^2 b|^2 + \frac{\kappa^2\lambda^2}{2}(1-|v|^2)^2 \\
&\geq \frac{J_{\kappa\lambda,B_{R/\lambda}}(v,B)}{\kappa^2R^2} + \frac{1}{2\kappa^2R^2}\left(\frac{1}{\lambda^2}-1\right)\int_{B_{R/\lambda}} |\operatorname{curl} B - \kappa^2\lambda^2 b|^2 \\
&\geq \frac{\min J_{\kappa\lambda,B_{R/\lambda}}}{(\kappa\lambda)^2(R/\lambda)^2} + \frac{1}{2\kappa^2R^2}(1-\lambda^2)\int_{B_R} |\operatorname{curl} A - b\kappa^2|^2.
\end{aligned}$$

But from Lemma 3.2, we have

$$\frac{\min J_{\kappa\lambda,B_{R/\lambda}}}{(\kappa\lambda)^2(R/\lambda)^2} = \frac{\min J_{\kappa\lambda,B_R}}{(\kappa\lambda)^2R^2} + o(1).$$

Hence, we deduce that for all $\lambda < 1$,

$$\frac{\min J_{\kappa,B_R}}{\kappa^2R^2} \geq \frac{\min J_{\kappa\lambda,B_R}}{(\kappa\lambda)^2R^2} + o(1) + (1-\lambda^2)\frac{1}{2\kappa^2R^2}\int_{B_R} |\operatorname{curl} A - b\kappa^2|^2. \tag{3.12}$$

Hence $\frac{\min J_{\kappa,B_R}}{\kappa^2R^2}$ is monotonic (up to $o(1)$) with respect to κ and must have a limit as $\kappa \to \infty$, which depends only on b. We denote it by $f(b)$. Then letting κ tend to infinity in (3.12) yields

$$f(b) \geq f(b) + \limsup_{\kappa\to\infty}(1-\lambda^2)\frac{1}{2\kappa^2R^2}\int_{B_R} |\operatorname{curl} A - b\kappa^2|^2;$$

thus we also deduce that

$$\frac{1}{2\kappa^2 R^2}\int_{B_R}|h-h_{\rm ex}|^2=o(1). \tag{3.13}$$

This means that, for energy-minimizers, the term $\int_\Omega |h-h_{\rm ex}|^2$ is negligible in the energy. This will be helpful later.

We now prove that f is monotonic. Let $\lambda<1$ again and let J_{b,B_R} now denote the functional restricted to B_R for the value $b\kappa^2$ of the applied field. Let us consider the same v and B as defined previously. By definition,

$$\begin{aligned}J_{\lambda^2 b,B_{R/\lambda}}(v,B)&=\frac12\int_{B_{R/\lambda}}|\nabla_B v|^2+|\operatorname{curl}B-\lambda^2 b\kappa^2|^2+\frac{\kappa^2}{2}(1-|v|^2)^2\\&=\frac{1}{\lambda^2}J_{b,B_R}(u,A)-\left(\frac{1}{\lambda^2}-1\right)\frac12\int_{B_{R/\lambda}}|\nabla_B v|^2-\left(\frac{1}{\lambda^4}-1\right)\frac12\int_{B_{R/\lambda}}|\operatorname{curl}B-\lambda^2 b\kappa^2|^2,\end{aligned}$$

where we have used (3.11). Thus, using (3.13),

$$J_{\lambda^2 b,B_{R/\lambda}}(v,B)=\frac{1}{\lambda^2}J_{b,B_R}(u,A)-\left(\frac{1}{\lambda^2}-1\right)\frac12\int_{B_R}|\nabla_A u|^2-o(1). \tag{3.14}$$

Therefore,

$$\frac{\min J_{\lambda^2 b,B_{R/\lambda}}}{\kappa^2(R/\lambda)^2}\le\frac{J_{b,B_R}(u,A)}{\lambda^2\kappa^2(R/\lambda)^2}.$$

In view of the previous results, the left-hand side of this inequality converges to $f(\lambda^2 b)$, while the right-hand side converges to $f(b)$. We deduce that for all $\lambda<1$,

$$f(\lambda^2 b)\le f(b);$$

thus f is nondecreasing. One can even deduce from (3.14) that f is increasing, because $\liminf \frac{1}{\kappa^2R^2}\int_{B_R}|\nabla_A u|^2>0$. Now taking $\lambda\ge1$, we get, as in (3.14), that

$$f(\lambda^2 b)\le f(b)-\psi(\lambda),$$

where $\psi(\lambda)\to0$ as $\lambda\to1$. This implies that f is continuous. □

In view of the result of Proposition 1.3, we have, for $b\le1$,

$$\frac{b}{2}-\frac{b^2}{4}\le f(b)\le\frac14. \tag{3.15}$$

We will prove the upper bound on f in the next section. Leaving it aside, let us now complete the proof of the theorem.

End of the proof of Theorem 1.4. Taking the scalar product of the first Ginzburg–Landau equation (1.2) with u yields the standard equation for $\rho=|u|$:

$$-\Delta\rho+\rho|\nabla\varphi-A|^2=\kappa^2\rho(1-\rho^2). \tag{3.16}$$

Then we multiply it by ρ and integrate. We are led, after integration by parts (using (1.5)), to

$$\int_\Omega|\nabla\rho|^2+\rho^2|\nabla\varphi-A|^2=\int_\Omega\kappa^2\rho^2(1-\rho^2).$$

We deduce the following relation, true for any solution of Ginzburg–Landau:

$$J(u,A)=\frac{\kappa^2}{4}\int_\Omega(1-\rho^4)+\frac{1}{2}\int_\Omega|h-h_{\rm ex}|^2. \tag{3.17}$$

In view of (3.13), if (u,A) is an energy-minimizer, this becomes

$$J(u,A)=\frac{\kappa^2}{4}\int_\Omega(1-\rho^4)+o(\kappa^2).$$

If we integrate over B_R instead of Ω and use (2.2) to handle the boundary term, we find, still for minimizers,

$$J_{B_R}(u,A)=\frac{\kappa^2}{4}\int_{B_R}(1-\rho^4)+O(\kappa R)+o(\kappa^2R^2). \tag{3.18}$$

Applying (3.18) to u and $\overline{u}$ successively gives (1.16).

Then

$$\frac{1}{|B_R|}\int_{B_R}|u|^4\to 1-4f(b)$$

for all $R\gg\varepsilon$, which implies the weaker conclusion that $|u|^4\rightharpoonup 1-4f(b)$ in L^∞ weak-*.

We also deduce from (1.14) combined with (1.19) that (1.20) holds. Plugging (1.16) in (1.20), we obtain

$$\begin{aligned}\frac{1}{|B_R|}\int_{B_R}|u|^2&\geq\frac{(1-b)^2+1-8f(b)+2b-b^2}{2(1-b)}\\&\geq\frac{1-4f(b)}{1-b}\geq\alpha(1-b),\end{aligned}$$

while

$$\frac{1}{|B_R|}\int_{B_R}|u|^2\leq\sqrt{\frac{1}{|B_R|}\int_{B_R}|u|^4}$$

comes from the Cauchy–Schwarz inequality.

This completes the proof of the theorem. □

4. Construction of test configurations. The upper bound of Theorem 1.4 relies on the construction of two test configurations, one being more interesting when $b\to 1$, the other one when $b\to 0$. Let us start with the first one, which follows somehow the construction of vortex solutions of [JT] in the self-dual situation.

PROPOSITION 4.1. *With the notation of the previous section, there exists a universal constant* $0<\alpha<1$ *such that*

$$f(b)\leq\frac{1-\alpha(1-b)_+^2}{4}. \tag{4.1}$$

Proof. Assume $b\leq 1$ (otherwise the conclusion is trivial). We construct a test configuration which is periodic with respect to a square lattice of size $\sqrt{\frac{2\pi}{b}}\varepsilon$. Let

$K_{\sqrt{\frac{2\pi}{b}}\varepsilon}$ denote an elementary square of the lattice and let $K_{\sqrt{2\pi}}$ denote the square of size $\sqrt{2\pi}$ centered at the origin. We solve for

$$\begin{cases} \Delta \log \rho_0 + 1 = 2\pi\delta_0 & \text{in } K_{\sqrt{2\pi}}, \\ \dfrac{\partial \log \rho_0}{\partial n} = 0 & \text{on } \partial K_{\sqrt{2\pi}}, \\ \displaystyle\int_{K_{\sqrt{2\pi}}} \log \rho_0 = 0. \end{cases} \tag{4.2}$$

There exists a (unique) solution to this system because the volume of $K_{\sqrt{2\pi}}$ is 2π. Let (r, θ) be the polar coordinates in the plane. We observe that $\log \rho_0 - \log r$ is smooth in $K_{\sqrt{2\pi}}$; hence $\frac{\rho}{r}$ too, and thus $\rho_0(0) = 0$. We then define in $K_{\sqrt{\frac{2\pi}{b}}\varepsilon}$,

$$\rho(x) = C\rho_0\left(\frac{x\sqrt{b}}{\varepsilon}\right),$$

where C minimizes $\int_{K_{\sqrt{\frac{2\pi}{b}}\varepsilon}} (1 - b - C^2\rho_0^2(\frac{x\sqrt{b}}{\varepsilon}))^2$, i.e. (after a little computation),

$$\rho(x) = \sqrt{1-b}\sqrt{\frac{\int_{K_{\sqrt{2\pi}}} \rho_0^2}{\int_{K_{\sqrt{2\pi}}} \rho_0^4}}\,\rho_0\left(\frac{x\sqrt{b}}{\varepsilon}\right). \tag{4.3}$$

One can see that ρ is a solution of

$$\begin{cases} \Delta \log \rho + \frac{b}{\varepsilon^2} = 2\pi\delta_0 & \text{in } K_{\sqrt{\frac{2\pi}{b}}\varepsilon}, \\ \dfrac{\partial \log \rho}{\partial n} = 0 & \text{on } \partial K_{\sqrt{\frac{2\pi}{b}}\varepsilon}. \end{cases} \tag{4.4}$$

ρ_0 is symmetric with respect to the axes of symmetry of the square and $\frac{\partial \rho_0}{\partial n} = 0$ on $\partial K_{\sqrt{2\pi}}$; thus we may extend ρ to any ball B_R ($R \gg \varepsilon$) by periodicity and get a C^1 function, which vanishes on a lattice Λ.

We then pick A to solve

$$\begin{cases} \operatorname{curl} A = b\kappa^2 & \text{in } B_R, \\ \operatorname{div} A = 0 & \text{in } B_R \end{cases}$$

and φ to satisfy

$$\nabla\varphi = \frac{\nabla^\perp \rho}{\rho} + A = \nabla^\perp \log \rho + A. \tag{4.5}$$

To achieve this, we fix a point x_0 of $B_R\backslash\Lambda$ and define

$$\varphi(x) = \int_{x_0}^{x} \partial_n \log \rho + A \cdot \tau.$$

This definition does not depend on the path joining x_0 to x, modulo 2π. Indeed, if $\gamma = \partial\omega$ is a closed path in $B_R\backslash\Lambda$ with positive orientation, using (4.4), we have

$$\int_\gamma \partial_n \log \rho + A \cdot \tau = \int_\omega \Delta \log \rho + \operatorname{curl} A = \int_\omega \Delta \log \rho + b\kappa^2 = 2\pi card(\omega \cap \Lambda) \in 2\pi\mathbb{Z}.$$

Hence $e^{i\varphi(x)}$ is well defined in $B_R\backslash\Lambda$. We then take

$$u(x) = \rho(x)e^{i\varphi(x)},$$

which has a continuous extension in B_R because ρ vanishes on Λ. Once this test configuration (u, A) is constructed, we evaluate its energy. In view of (2.9),

$$J_{B_R}(u, A) = |B_R|\kappa^2\left(\frac{b}{2} - \frac{b^2}{4}\right) + \frac{1}{2}\int_{B_R} |\mathcal{D}_A u|^2 + \frac{\kappa^2}{2}(1 - b - \rho^2)^2 + O(R\kappa^2). \tag{4.6}$$

But $|\mathcal{D}_A u|^2 = |\rho(\nabla\varphi - A) - \nabla^\perp\rho|^2 = 0$ by construction; cf. (4.5). Moreover, from (4.3),

$$\begin{aligned}\int_{K_{\sqrt{\frac{2\pi}{b}}\varepsilon}} (1 - b - \rho^2)^2 &= (1 - b)^2 \int_{K_{\sqrt{\frac{2\pi}{b}}\varepsilon}} \left(1 - \frac{\int_{K_{\sqrt{2\pi}}} \rho_0^2}{\int_{K_{\sqrt{2\pi}}} \rho_0^4}\, \rho_0^2\left(\frac{x\sqrt{b}}{\varepsilon}\right)\right) dx \\ &= (1 - b)^2\left(\frac{2\pi}{b}\varepsilon^2 - \frac{\varepsilon^2}{b}\frac{\left(\int_{K_{\sqrt{2\pi}}} \rho_0^2\right)^2}{\int_{K_{\sqrt{2\pi}}} \rho_0^4}\right). \qquad (4.7)\end{aligned}$$

Let us write

$$\alpha = \frac{\left(\int_{K_{\sqrt{2\pi}}} \rho_0^2\right)^2}{2\pi\int_{K_{\sqrt{2\pi}}} \rho_0^4}.$$

Since ρ_0 is not a constant function (see (4.2)), we have a strict Cauchy–Schwarz inequality

$$\left(\int_{K_{\sqrt{2\pi}}} \rho_0^2\right)^2 < 2\pi\int_{K_{\sqrt{2\pi}}} \rho_0^4,$$

and hence $0 < \alpha < 1$. Then, from (4.7),

$$\int_{B_R} (1 - b - \rho^2)^2 = |B_R|(1 - b)^2(1 - \alpha) + o(R^2). \tag{4.8}$$

Combining (4.6) and (4.8), we are led to

$$J_{B_R}(u, A) = |B_R|\kappa^2\left(\frac{b}{2} - \frac{b^2}{4} + \frac{(1 - b)^2(1 - \alpha)}{4}\right) + o(\kappa^2 R^2).$$

We conclude that

$$f(b) \le \limsup_{\kappa\to\infty} \frac{\min J_{B_R}}{\kappa^2|B_R|} \le \limsup_{\kappa\to\infty} \frac{J_{B_R}(u, A)}{\kappa^2|B_R|} \le \frac{1 - \alpha(1 - b)^2}{4}. \qquad \square$$

PROPOSITION 4.2. *There exists a universal constant c such that for $b \le 1$,*

$$f(b) \le \frac{b}{4}\left(\log\frac{1}{b} + c\right). \tag{4.9}$$

Proof. This estimate is stronger than (4.1) when $b \to 0$ and corresponds to a regime in which the distance between vortices is rather large compared to their core size ε, i.e., is close to the regime described in [SS2]. In order to prove this estimate, we just adjust the construction of a test function that we did in [SS2].

This test function is again periodic with respect to a square lattice of size $\sqrt{\frac{2\pi}{b}}\varepsilon$. Let us consider an elementary square $K_{\sqrt{\frac{2\pi}{b}}\varepsilon}$ centered at the origin, with B_ε the ball of radius ε centered at the origin, which is included in $K_{\sqrt{\frac{2\pi}{b}}\varepsilon}$ for all $b \leq 1$. The centers of the squares of the lattice will be denoted a_i. We take a $\rho \leq 1$, which satisfies

$$\begin{cases} \rho \equiv 1 & \text{in } K_{\sqrt{\frac{2\pi}{b}}\varepsilon} \backslash B_\varepsilon, \\ \rho \equiv 0 & \text{in } B_{\varepsilon/2}, \\ \displaystyle\int_{K_{\sqrt{\frac{2\pi}{b}}\varepsilon}} |\nabla\rho|^2 + \frac{\kappa^2}{2}(1-\rho^2)^2 \leq C. \end{cases} \tag{4.10}$$

Then we take h such that

$$\begin{cases} -\Delta h + h = \frac{8}{\varepsilon^2}\mathbf{1}_{B_{\varepsilon/2}} & \text{in } K_{\sqrt{\frac{2\pi}{b}}\varepsilon}, \\ \dfrac{\partial h}{\partial n} = 0 & \text{on } \partial K_{\sqrt{\frac{2\pi}{b}}\varepsilon}, \end{cases} \tag{4.11}$$

where $\mathbf{1}$ denotes a characteristic function. We extend ρ and h by periodicity to B_R ($R \gg \varepsilon$) and pick A such that $\operatorname{curl} A = h$ and $\operatorname{div} A = 0$. Then we take φ such that

$$\nabla\varphi = -\nabla^\perp h + A, \tag{4.12}$$

i.e., by choosing a point x_0 in $B_R \backslash \cup_i B_{\varepsilon/2}(a_i)$ and setting

$$\varphi(x) = \int_{x_0}^{x} -\frac{\partial h}{\partial n} + A \cdot \tau.$$

This integral does not depend on the path joining x_0 to x in $B_R \backslash \cup_i B_{\varepsilon/2}(a_i)$, modulo 2π. This can be seen from (4.11). Thus $e^{i\varphi}$ is well defined in $B_R \backslash \cup_i B_{\varepsilon/2}(a_i)$, and

$$u(x) = \rho(x) e^{i\varphi(x)}$$

has a meaning on all of B_R (since $\rho \equiv 0$ in $\cup_i B_{\varepsilon/2}(a_i)$). Exactly as in [SS2], one shows that

$$\frac{1}{2}\int_{K_{\sqrt{\frac{2\pi}{b}}\varepsilon}} |\nabla h|^2 + |h - h_{\text{ex}}|^2 \leq \pi \log \frac{\sqrt{\frac{1}{b}}\varepsilon}{b} + C = \frac{\pi}{2}\log\frac{1}{b} + C. \tag{4.13}$$

There remains the matter of evaluating the energy of (u, A) per square. From (4.12), we have $\rho^2|\nabla\varphi - A|^2 \leq |\nabla h|^2$, and hence

$$\begin{aligned} J_K(u,A) &= \int_{K_{\sqrt{\frac{2\pi}{b}}\varepsilon}} |\nabla\rho|^2 + \rho^2|\nabla\varphi - A|^2 + |h - h_{\text{ex}}|^2 + \frac{\kappa^2}{2}(1-\rho^2)^2 \\ &\leq \frac{\pi}{2}\log\frac{1}{b} + c. \end{aligned}$$

Multiplying this estimate by the number of squares in B_R, $\frac{|B_R|b}{2\pi\varepsilon^2}$, we find

$$J_{B_R}(u,A) \leq |B_R|\kappa^2 \left(\frac{b}{4}\log\frac{1}{b} + cb\right).$$

We then conclude, as in the previous proposition, that (4.9) holds. ∎

REFERENCES

[AD] A. Aftalion and N. Dancer, *On the symmetry and uniqueness of solutions of the Ginzburg–Landau equations for small domains*, Commun. Contemp. Math., 3 (2001), pp. 1–14.

[Al1] Y. Almog, *Non-linear surface superconductivity for type* II *superconductors in the large domain limit*, Arch. Ration. Mech. Anal., 165 (2002), pp. 271–293.

[Al2] Y. Almog, *On the bifurcation and stability of periodic solutions to the Ginzburg–Landau equations in the plane*, SIAM J. Appl. Math., 61 (2000), pp. 149–171.

[BBH] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg–Landau Vortices*, Birkhäuser Boston, Boston, MA, 1994.

[BPT] P. Bauman, D. Phillips, and Q. Tang, *Stable nucleation for the Ginzburg–Landau system with an applied magnetic field*, Arch. Ration. Mech. Anal., 142 (1998), pp. 1–43.

[BS] A. Bernoff and P. Sternberg, *Onset of superconductivity in decreasing fields for general domains*, J. Math. Phys., 39 (1998), pp. 1272–1284.

[C] S. J. Chapman, *Nucleation of superconductivity in decreasing fields*, European J. Appl. Math., 5 (1994), pp. 449–494.

[DeG] P.-G. De Gennes, *Superconductivity of Metal and Alloys*, Benjamin, New York, Amsterdam, 1966.

[DFS] M. Del Pino, P. Felmer, and P. Sternberg, *Boundary concentration for eigenvalue problems related to the onset of superconductivity*, Comm. Math. Phys., 210 (2000), pp. 413–446.

[Du] M. Dutour, *Bifurcation vers l'état d'Abrikosov et Diagramme de Phase*, Ph.D. thesis, University of Orsay, France, 1999.

[GP] T. Giorgi and D. Phillips, *The breakdown of superconductivity due to strong fields for the Ginzburg–Landau model*, SIAM J. Math. Anal., 30 (1999), pp. 341–359.

[HM] B. Helffer and A. Morame, *Magnetic bottles in connection with superconductivity*, J. Funct. Anal., 185 (2001), pp. 604–680.

[HP] B. Helffer and X. B. Pan, *Upper critical field and location of surface nucleation of superconductivity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[JT] A. Jaffe and C. Taubes, *Vortices and Monopoles*, Birkhäuser Boston, Boston, MA, 1980.

[LP1] K. Lu and X. B. Pan, *Gauge-invariant eigenvalue problems in* $\mathbb{R}^2$ *and* $\mathbb{R}^2_+$, Trans. Amer. Math. Soc., 352 (2000), pp. 1247–1276.

[LP2] K. Lu and X. B. Pan, *Eigenvalue problems of Ginzburg–Landau operator in bounded domains*, J. Math. Phys., 40 (1999), pp. 2647–2670.

[LP3] K. Lu and X. B. Pan, *Estimates of the upper critical field for the Ginzburg–Landau equations of superconductivity*, Phys. D, 127 (1999), pp. 73–104.

[LP4] K. Lu and X. B. Pan, *Surface nucleation of superconductivity in* 3*-dimensions*, J. Differential Equations, 168 (2000), pp. 386–452.

[M] R. Montgomery, *Hearing the zero locus of a magnetic field*, Comm. Math. Phys., 168 (1995), pp. 651–675.

[P1] X. B. Pan, *Surface superconductivity in applied magnetic fields above* H_{c_2}, Comm. Math. Phys., 228 (2002), pp. 327–370.

[P2] X. B. Pan, *Upper critical fields for superconductors with edges and corners*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 447–482.

[SdG] D. Saint-James and P. G. De Gennes, *Onset of superconductivity in decreasing fields*, Phys. Lett., 6 (1963), pp. 306–308.

[SST] D. Saint-James, G. Sarma, and E. Thomas, *Type-*II *Superconductivity*, Pergamon Press, Oxford, UK, 1969.

[SS1] E. Sandier and S. Serfaty, *Global minimizers for the Ginzburg–Landau functional below the first critical magnetic field*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 119–145.

[SS2] E. Sandier and S. Serfaty, *On the energy of type-*II *superconductors in the mixed phase*, Rev. Math. Phys., 12 (2000), pp. 1219–1257.

[SS3] E. Sandier and S. Serfaty, *A rigorous derivation of a free-boundary problem arising in superconductivity*, Ann. Sci. École Norm. Sup. (4), 33 (2000), pp. 561–592.

[SS4] E. Sandier and S. Serfaty, *Limiting vorticities for the Ginzburg–Landau equations*, Duke Math. J., to appear.

[SS5] E. Sandier and S. Serfaty, *Ginzburg–Landau minimizers near the first critical field have bounded vorticity*, Calc. Var. Partial Differential Equations, to appear.

[S1] S. Serfaty, *Local minimizers for the Ginzburg–Landau energy near critical magnetic field, part* I, Commun. Contemp. Math., 1 (1999), pp. 213–254.

[S2] S. Serfaty, *Local minimizers for the Ginzburg–Landau energy near critical magnetic field, part* II, Commun. Contemp. Math., 1 (1999), pp. 295–333.

[S3] S. Serfaty, *Stable configurations in superconductivity: Uniqueness, multiplicity and vortex-nucleation*, Arch. Ration. Mech. Anal., 149 (1999), pp. 329–365.

[T] M. Tinkham, *Introduction to Superconductivity*, 2nd ed., McGraw–Hill, New York, 1996.

SIAM J. MATH. ANAL.
Vol. 34, No. 4, pp. 957–991

SUPERCONDUCTING FILMS IN PERPENDICULAR FIELDS AND THE EFFECT OF THE DE GENNES PARAMETER*

XING-BIN PAN†

Abstract. In this paper we study superconductivity of a thin film placed in a perpendicular magnetic field. We discuss the dependence of the upper critical field H_{C_3} on the thickness l of the film and the Ginzburg–Landau parameter κ, and we examine nucleation of superconductivity. We show that a critical change occurs at $l = 2\gamma\kappa^{-2}$. If $l > a\kappa^{-2}$ $(a > 2\gamma)$, the film exhibits type II behaviors: as the applied magnetic field decreases from H_{C_3}, superconductivity nucleates in a strip at the lateral surface and develops a lateral surface superconducting state. If $l \le 2\gamma\kappa^{-2} + C\kappa^{-4}$, the film exhibits type I behaviors.

Key words. superconductivity, Ginzburg–Landau system, critical field, Schrödinger operator with a magnetic field, de Gennes parameter, thin film

AMS subject classifications. 35Q55, 82D55

PII. S0036141002406734

1. Introduction.

Motivated by the recent work of Richardson and Rubinstein [RR1, RR2], we study the effect of the de Gennes parameter on superconductivity of thin films.

Mathematical models of superconducting thin films have been studied recently by many authors; see, for instance, Du and Gunzburger [DG], Chapman, Du, and Gunzburger [CDG], Chen, Elliott, and Tang [CET], Berger and Rubinstein [BR1, BR2], Rubinstein and Schatzman [RS], Richardson and Rubinstein [RR1, RR2], Jimbo and Morita [JM], Ding and Du [DD], and references therein.[1] Among other things, it was shown in [CDG] that all superconducting materials, whether type I (with small Ginzburg–Landau parameter κ) or type II (with large κ), behave as type II superconductors when made into sufficiently thin films; and for a very thin film placed in a magnetic field, only the perpendicular component of the applied field has influence on the superconductivity.[2] On the other hand, the recent works of Richardson and Rubinstein [RR1, RR2] show that the de Gennes parameter has an important effect on superconductivity of thin films. It was mentioned in [RR1] that "the effect of the de Gennes boundary condition is to depress the temperature at which superconductivity occurs," and they conjectured that "for sufficiently thin wires or small de Gennes distance,[3] superconductivity may never be favorable." This question motivated us to study nucleation of superconductivity of thin films in a general content.

Let us consider a superconducting film of thickness l and a cross-section Ω:

$$D_l = \{(x, z) : \ x \in \Omega, \ 0 < z < l\},$$

*Received by the editors May 2, 2002; accepted for publication (in revised form) October 31, 2002; published electronically April 9, 2003. This work was partially supported by the National Natural Science Foundation of China, the Science Foundation of the Ministry of Education of China, the Zhejiang Provincial Natural Science Foundation of China, and NUS Academic Research grants R-146-000-022-112 and R-146-000-033-112.

http://www.siam.org/journals/sima/34-4/40673.html

†Department of Mathematics, National University of Singapore, Singapore 119260 (matpanxb@nus.edu.sg), and Department of Mathematics, Zhejiang University, Hangzhou 310027, China.

[1] Superconductivity on samples with small size has also been studied by Aftalion and Dancer [ADa] and Aftalion and Du [ADu].

[2] See also [HT, Chapter 8].

[3] A large de Gennes parameter gives a small de Gennes distance.

where $0 < l \ll 1$, and Ω is a bounded, simply-connected, and smooth (C^4) domain in $\mathbb{R}^2$. Throughout this paper, $x = (x_1, x_2)$ denotes a point in $\bar{\Omega}$, (x, z) denotes a point in D_l; ds denotes the measure on $\partial\Omega$, dS denotes the measure on ∂D_l; ν denotes the unit outer-normal vector of $\partial\Omega$, and ν_D denotes the unit outer-normal vector of the boundary of D_l, which is well defined in $\partial^* D_l$, where

$$\partial^* D_l = \partial\Omega \times (0, l) \bigcup \Omega \times \{0, l\}.$$

Let

$$d_{\partial\Omega}(x) = \operatorname{dist}(x, \partial\Omega), \qquad d_l(x, z) = \operatorname{dist}((x, z), \partial D_l).$$

Let κ_r be the curvature of $\partial\Omega$, and set

$$\kappa_{\max} = \max_{x\in\partial\Omega} \kappa_r(x), \qquad \mathcal{N}(\partial\Omega) = \{x \in \partial\Omega : \ \kappa_r(x) = \kappa_{\max}\}.$$

According to the Ginzburg–Landau theory, superconductivity is described by a complex-valued function ψ (order parameter) and a real vector field $\mathcal{A}$ (magnetic potential), and $(\psi, \mathcal{A})$ is a minimizer of the Ginzburg–Landau functional. Let us consider a homogeneous applied magnetic field $\mathcal{H} = \sigma\mathbf{h}$, where $\mathbf{h}$ is a constant unit vector and σ is a positive number. In this paper, as we are concerned with the behavior of a film in a perpendicular field, $\mathbf{h}$ is chosen to be perpendicular to Ω:

$$\mathbf{h} = \mathbf{e}_3 = (0, 0, 1).$$

We shall treat σ as a parameter. So we set $\mathcal{A} = \sigma\mathbf{A}$. With a proper scaling, we may rewrite the Ginzburg–Landau functional as (see [GL, dG, CHO, DGP, R])

$$\begin{aligned}\mathcal{G}[\psi, \mathbf{A}] = \int_{D_l} \left\{ |\nabla_{\kappa\sigma\mathbf{A}}\psi|^2 + \frac{\kappa^2}{2}(|\psi|^2 - 1)^2 \right\} dxdz + \gamma \int_{\partial D_l} |\psi|^2 dS \\ + \kappa^2\sigma^2 \int_{\mathbb{R}^3} |\operatorname{curl} \mathbf{A} - \mathbf{h}|^2 dxdz.\end{aligned}$$

The minimizers $(\psi, \mathbf{A})$ satisfy the following Ginzburg–Landau system:

(1.1)
$$\begin{cases} -\nabla^2_{\kappa\sigma\mathbf{A}}\psi = \kappa^2(1 - |\psi|^2)\psi & \text{in } D_l, \\ \operatorname{curl}^2\mathbf{A} = -\dfrac{i}{2\kappa\sigma}(\bar{\psi}\nabla\psi - \psi\nabla\bar{\psi}) - |\psi|^2\mathbf{A} & \text{in } D_l, \\ \operatorname{curl}^2\mathbf{A} = \mathbf{0} & \text{in } \mathbb{R}^3 \setminus D_l, \\ (\nabla_{\kappa\sigma\mathbf{A}}\psi)\cdot\nu_D + \gamma\psi = 0, \quad [\nu_D \cdot \mathbf{A}] = 0, \quad [\nu_D \times \operatorname{curl}\mathbf{A}] = \mathbf{0} & \text{on } \partial^* D_l, \\ \operatorname{curl}\mathbf{A} \to \mathbf{h} & \text{as } |(x, z)| \to \infty. \end{cases}$$

Here $i = \sqrt{-1}$, κ is the Ginzburg–Landau parameter, and $[\cdot]$ denotes the jump in the enclosed quantity across ∂D_l. The boundary condition for ψ in the fourth equality was posed by de Gennes [dG] for a superconductor adjacent to other material. $\gamma \geq 0$ is the de Gennes parameter.[4] γ is very small for insulators, very large for magnetic

[4] In the literature $d = 1/\gamma$ is called the de Gennes distance.

materials, and lying in between for nonmagnetic materials. In this paper we shall always assume that $\gamma > 0$. Note that in the above system the unit of length is the penetration depth.

We call a minimizer $(\psi, \mathbf{A})$ of the functional $\mathcal{G}$ a *minimal solution* of (1.1). Let $\mathbf{F_h}$ be a smooth vector field defined on $\mathbb{R}^3$ such that

$$\text{curl } \mathbf{F_h} = \mathbf{h} \quad \text{and} \quad \text{div } \mathbf{F_h} = 0 \quad \text{in } \mathbb{R}^3. \tag{1.2}$$

$(0, \mathbf{F_h})$ is a trivial solution of (1.1), and it is the only minimal solution if σ is large. As in [LP4], we define the upper critical field (here emphasizing the dependence of H_{C_3} on the thickness of the film) by

$$H_{C_3}(\mathbf{h}, \kappa, l) = \inf\{\sigma > 0 : (0, \mathbf{F_h}) \text{ is a minimizer of } \mathcal{G}\}.$$

We are interested in the dependence of H_{C_3} on κ, l, γ and on the geometry of the cross-section of the film. Using the methods developed in [LP1, LP2, LP3, LP4, LP5, LP6, HM1, HP], and especially in [P1], we are able to establish an estimate of $H_{C_3}(\mathbf{h}, \kappa, l)$ for small l.

THEOREM 1.1. *Let* $\mathbf{h}$ *be a unit vector perpendicular to* Ω *and let* $\gamma > 0$ *be given. Let* a*,* b*, and* c *be fixed positive constants. For large* κ *we have*

(1.3)
$$H_{C_3}(\mathbf{h}, \kappa, l) = \begin{cases} \frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma) - \frac{2\gamma}{a\beta_0} + o(1) & \text{if } l = a\kappa^{-1},\ a > 0, \\ (1 - \frac{2\gamma}{a})\frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma)(1 - \frac{2\gamma}{a})^{1/2} + o(1) & \text{if } l = a\kappa^{-2},\ a > 2\gamma, \\ \frac{c}{2\gamma\beta_0} + O(\kappa^{-1}) & \text{if } l = 2\gamma\kappa^{-2} + c\kappa^{-3},\ c > 0, \\ O(\kappa^{-1}) & \text{if } l = 2\gamma\kappa^{-2} + b\kappa^{-4},\ b \geq b_0, \\ 0 & \text{if } l = 2\gamma\kappa^{-2} + b\kappa^{-4},\ b < b_0, \end{cases}$$

where C_1 *and* β_0 *are universal constants and* b_0 *depends only on* Ω *and* γ.

The numbers C_1, β_0, and b_0 will be given in section 2. In section 6 we shall give a proof of Theorem 1.1 and also discuss nucleation of superconductivity.

Remark 1. One may expect that a superconducting thin film placed in a perpendicular magnetic field will behave as a two-dimensional superconductor.[5] In fact, this is true if $l > a\kappa^{-2}$ for some $a > 2\gamma$. Recall that for a two-dimensional superconductor Ω placed in an applied magnetic field perpendicular to Ω, we have the following (see [LP4, HP, P2]):

(a) For large κ,

$$H_{C_3}(\kappa) = \frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}\kappa_{\max} + O(\kappa^{-1/3}). \tag{1.4}$$

(b) As the applied field decreases from $H_{C_3}(\kappa)$, superconductivity nucleates first in the set of the maximum points of the curvature, $\mathcal{N}(\partial\Omega)$, and then develops a *surface superconducting state*.[6]

[5] By a *two-dimensional superconductor* we actually mean a superconducting cylinder with infinite height and a cross-section Ω, and placed in an applied magnetic field perpendicular to Ω. In this case one may reduce the Ginzburg–Landau system to a two-dimensional system on Ω.

[6] See [P2, footnote on p. 328] for the meaning of a "surface superconducting state."

We may call such types of behavior *two-dimensional type* II *behaviors*.[7] For a thin film $D_l = \Omega \times (0, l)$ with $l \geq a\kappa^{-2}$ $(a > 2\gamma)$, we have the following (see Theorem 1.1 above and Theorems 6.2 and 6.3 in section 6):

(a′) $H_{C_3}(\mathbf{h}, \kappa, l)$ has order of κ and depends on the curvature of the cross-section.

(b′) As the applied field decreases from H_{C_3}, superconductivity nucleates first in the strip $\mathcal{N}(\partial\Omega) \times [0, l]$ and then develops a *lateral surface superconducting state*.[8] So we can say that a film with thickness $l \geq a\kappa^{-2}$ $(a > 2\gamma)$ also exhibits two-dimensional type II behaviors when placed in a perpendicular field.

Remark 2. Equation (1.3) also indicates a thin film behavior, which is presented by the quantities involving the de Gennes parameter γ. These quantities become important when l is close to $2\gamma\kappa^{-2}$. In fact, $l = 2\gamma\kappa^{-2}$ is a critical value in the following sense:

(i) When $l \geq a\kappa^{-2}$ $(a > 2\gamma)$, the film exhibits a two-dimensional behavior of type II superconductors.

(ii) When $l \sim 2\gamma\kappa^{-2}$, the film behaves like a type I superconductor.

(iii) The film loses superconductivity if $l < 2\gamma\kappa^{-2} + b_0\kappa^{-4}$.

In order to estimate the value of H_{C_3}, we need an estimate of the lowest eigenvalue $\mu_\gamma(\mathbf{A})$ of the following problem associated with a given vector field $\mathbf{A}$:

$$\begin{cases} -\nabla^2_{\mathbf{A}}\phi = \mu\phi & \text{in } D_l, \\ (\nabla_{\mathbf{A}}\phi)\cdot\nu_D + \gamma\phi = 0 & \text{on } \partial^* D_l. \end{cases} \tag{1.5}$$

We especially need to estimate $\mu_\gamma(b\mathbf{F}_{\mathbf{h}})$ for large b. We shall discuss only a perpendicular field ($\mathbf{h} = (0, 0, 1)$) in this paper and wish to consider applied fields in general directions in the near future. We expect that the parallel components of the applied field play roles in determining the location of nucleation.

The outline of this paper is the following. In section 2 we collect some preliminary results that will be used in later sections. In section 3 we study an eigenvalue problem in a two-dimensional domain and extend the Helffer–Morame estimate [HM1] of the lowest eigenvalue to the problems with de Gennes boundary conditions. In section 4 we give some elliptic estimates for the minimizers of the Ginzburg–Landau functional in the films D_l, with constants independent of l. In section 5 we present estimates of the lowest eigenvalue for eigenvalue problems on the films. In section 6 we study superconductivity of the films in perpendicular magnetic fields, establish an estimate of H_{C_3}, and find the location of nucleation. From the results established in section 6 we get Theorem 1.1.

We should mention that the Ginzburg–Landau system with de Gennes boundary conditions has been studied in Lu and Pan [LP1, LP2, LP3, LP4, LP5, LP6]. In particular, Lu and Pan [LP1] discussed the Ginzburg–Landau system without applied fields and described the asymptotic behavior of the minimal solutions for large value of the de Gennes parameter. Combining the results in this paper and those in [LP1] we see that the effect of the de Gennes parameter is important when its value is large compared with the scale of the samples.

The upper critical field H_{C_3} and surface superconductivity have been studied by many physicists and mathematicians. For early research see Saint-James and de

[7] More precisely, two-dimensional behaviors are those of a cylindrical superconductor of infinite height and a constant cross-section in response to an applied magnetic field perpendicular to the cross-section.

[8] For a film, a *lateral surface superconducting state* is such a state that superconductivity is confined in a thin layer around the lateral surface, and superconductivity in the layer is not weak.

Gennes [SdG], Saint-James and Sarma [SST], and Tinkham [T]; for recent mathematical research on the estimates of H_{C_3} and the study of the surface nucleation phenomenon, see Chapman [C], Bauman, Phillips, and Tang [BPT], Giorgi and Phillips [GP], Bernoff and Sternberg [BS], Lu and Pan [LP1, LP2, LP3, LP4, LP5], del Pino, Felmer, and Sternberg [DFS], Jadallah [J], Pan [P1, P2, P3], Pan and Kwek [PK], Helffer and Morame [HM1, HM2], Helffer and Pan [HP], and Almog [Al].

2. Preliminaries. In this section we give some preliminary results which will be used in later sections. We first recall an eigenvalue variation problem. For every constant z, let $\beta(z)$ denote the lowest eigenvalue of the following problem in $L^2(\mathbb{R}_+)$:

$$-u'' + (z+t)^2 u = \beta(z)u \quad \text{for } t > 0, \qquad u'(0) = 0. \tag{2.1}$$

It was first proved by Dauge and Helffer [DH] (also see Bolley and Helffer [BH]) that[9] there is a unique z_0, $z_0 < 0$ such that

$$\beta_0 \equiv \beta(z_0) = \inf_{z\in\mathbb{R}} \beta(z). \tag{2.2}$$

Moreover, $\beta_0 = z_0^2$ and $0.5 < \beta_0 < 0.76$. Throughout this paper, β_0 always denotes the number given in (2.2), and C_1 always denotes the number defined by

$$C_1 = \frac{u^2(0)}{3\|u\|^2_{L^2(\mathbb{R}_+)}},$$

where $u(t)$ is the positive eigenfunction of (2.1) for $z = z_0$ and $\beta = \beta_0$. β_0 and C_1 appeared in (1.3) and (1.4) and will be used frequently in later sections.

In Lemma 2.1 below we shall give a simple estimate for the lowest eigenvalue $\mu_\gamma(\mathbf{A})$ of (1.5). For this purpose, we need the numbers $\beta_{\gamma,\Omega}$, $\mu_{\gamma,\Omega}$, and $\tau_\gamma(l)$, where

$$\beta_{\gamma,\Omega} = \inf_{\phi\in W^{1,2}(\Omega)} \frac{\int_\Omega |\nabla\phi|^2 dx + \gamma\int_{\partial\Omega}|\phi|^2 ds}{\int_\Omega |\phi|^2 dx},$$
$$\mu_{\gamma,D_l} = \inf_{\phi\in W^{1,2}(D_l)} \frac{\int_{D_l} |\nabla\phi|^2 dxdz + \gamma\int_{\partial D_l}|\phi|^2 dS}{\int_{D_l} |\phi|^2 dxdz},$$

and $\lambda = \tau_\gamma(l)^2$ is the lowest eigenvalue of

$$\begin{cases} -\xi'' = \lambda\xi & \text{for } 0 < t < l, \\ \xi'(0) = \gamma\xi(0), \qquad \xi'(l) = -\gamma\xi(l). \end{cases} \tag{2.3}$$

$\tau_\gamma(l)$ is determined by the smallest positive solution of the algebraic equation

$$\tan(\tau_\gamma(l)l) = \frac{2\gamma\tau_\gamma(l)}{\tau_\gamma(l)^2 - \gamma^2},$$

and the eigenfunctions of (2.3) associated with $\tau_\gamma(l)^2$ are $c\xi_l(t)$, where

$$\xi_l(t) = \tau_\gamma(l)\cos(\tau_\gamma(l)t) + \gamma\sin(\tau_\gamma(l)t). \tag{2.4}$$

[9]It was proved again by Lu and Pan [LP2] (also see [LP4]) without their knowing the results of [DH] and [BH]. The methods are different though.

When $0 < l < \frac{\pi}{2\gamma}$, we have $\tau_\gamma(l) > \gamma$, and $\tau_\gamma(l)l \to 0$ as $l \to 0$. Using Taylor expansion we find

$$\frac{2\gamma}{l} - \frac{\gamma^2}{3} < \tau_\gamma(l)^2 < \frac{2\gamma}{l} - \frac{\gamma^2}{3} + O(\gamma^3 l) \quad \text{as } l \to 0. \tag{2.5}$$

The following equality shows the relation between these numbers:

$$\mu_{\gamma,D_l} = \beta_{\gamma,\Omega} + \tau_\gamma(l)^2, \tag{2.6}$$

which is obtained by solving, using the separable variables method, the eigenvalue problem corresponding with μ_{γ,D_l}.

LEMMA 2.1. (i) *For any vector field* $\mathbf{A}$*, the lowest eigenvalue* $\mu_\gamma(\mathbf{A})$ *of* (1.5) *satisfies*

$$\mu_\gamma(\mathbf{A}) \geq \mu_{\gamma,D_l} > \beta_{\gamma,\Omega} + \frac{2\gamma}{l} - \frac{\gamma^2}{3}.$$

(ii) *For any unit vector* $\mathbf{h}$*,* $H_{C_3}(\mathbf{h}, \kappa, l) > 0$ *if and only if* $\mu_{\gamma,D_l} < \kappa^2$.

Proof. For any vector field $\mathbf{A}$ and any $\phi \in W^{1,2}(D_l)$, from Kato's inequality we have

$$\int_{D_l} |\nabla_{\mathbf{A}}\phi|^2 dxdz \geq \int_{D_l} \big|\nabla|\phi|\big|^2 dxdz.$$

Hence $\mu_\gamma(\mathbf{A}) \geq \mu_{\gamma,D_l}$. We get (i) from this and (2.5), (2.6).

If $H_{C_3}(\mathbf{h}, \kappa, l) > 0$, then for $0 < \sigma < H_{C_3}$, the Ginzburg–Landau functional $\mathcal{G}$ has a nontrivial minimizer $(\psi, \mathbf{A})$ and thus the lowest eigenvalue $\mu_\gamma(\sigma\kappa\mathbf{A}) < \kappa^2$; see [LP4, Lemma 2.1]. From conclusion (i) we must have $\mu_{\gamma,D_l} < \kappa^2$.

On the other hand, if $\mu_{\gamma,D_l} < \kappa^2$, then for the vector field $\mathbf{F_h}$ given in (1.2), $\mu_\gamma(\sigma\kappa\mathbf{F_h}) < \kappa^2$ for all sufficiently small $\sigma > 0$. Hence the functional $\mathcal{G}$ has a nontrivial minimizer; see [LP4, Lemma 2.1]. Thus $H_{C_3}(\mathbf{h}, \kappa, l) > 0$. (ii) is proved. □

Remark 3. (i) As a consequence of Lemma 2.1, if $\kappa^2 + \frac{\gamma^2}{3} - \beta_{\gamma,\Omega} > 0$ and if

$$0 < l < \frac{2\gamma}{\kappa^2 + \frac{\gamma^2}{3} - \beta_{\gamma,\Omega}}, \tag{2.7}$$

then $\mu_{\gamma,D_l} > \kappa^2$, and hence $\mu_\gamma(\mathbf{A}) > \kappa^2$ for any vector field $\mathbf{A}$. Thus (1.1) has no nontrivial solution; see [LP4, Lemma 2.1]. Hence superconductivity is not favorable for a very thin film with its thickness l satisfying (2.7). This verifies the prediction of Richardson and Rubinstein [RR1, RR2].

(ii) Let us define $l = l(\kappa)$ to be the positive root of the equation

$$\mu_{\gamma,D_l} = \kappa^2.$$

Lemma 2.1 indicates that nontrivial minimal solutions $(\psi, \mathbf{A})$ of the Ginzburg–Landau system (1.1) bifurcate from $(0, \mathbf{0})$ when the parameter (κ, l) moves upward in the parameter space $\{(\kappa, l) : \ \kappa > 0, \ l > 0\}$ away from the bifurcation point $(\kappa, l(\kappa))$. From (2.5) we find that for large κ,

$$l(\kappa) = 2\gamma\kappa^{-2} + 2\gamma\left(\beta_{\gamma,\Omega} - \frac{\gamma^2}{3}\right)\kappa^{-4} + O(\kappa^{-6}).$$

(iii) The number b_0 that appeared in (1.3) is defined by

$$b_0 = 2\gamma\left(\beta_{\gamma,\Omega} - \frac{\gamma^2}{3}\right). \tag{2.8}$$

If $l = 2\gamma\kappa^{-2} + b\kappa^{-4}$ with $b < b_0$ and κ is large, the minimizers of the Ginzburg–Landau functional $\mathcal{G}$ are trivial. This verifies the last conclusion in Theorem 1.1.

(iv) Since superconducting states exist on thin films only when $l > l(\kappa)$, we assume in the following that

$$\kappa^2 > \mu_{\gamma,D_l}.$$

It implies that $\kappa \to \infty$ as $l \to 0$. Thus we consider only thin films with large value of κ. We remark that a bulk superconductor is type II if κ is large; however, a thin film with large κ and $l \sim 2\gamma\kappa^{-2}$ may present a type I behavior; see Theorem 6.5 below.

3. A two-dimensional eigenvalue problem. In this section we estimate the lowest eigenvalue $\beta_\gamma(\varepsilon^{-2}\mathbf{A})$, as $\varepsilon \to 0$, of the problem

$$\begin{cases} -\nabla^2_{\varepsilon^{-2}\mathbf{A}}\phi = \beta\phi & \text{in } \Omega, \\ (\nabla_{\varepsilon^{-2}\mathbf{A}}\phi)\cdot\nu + \gamma\phi = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.1}$$

where Ω is a bounded C^4 domain in $\mathbb{R}^2$, γ is a positive constant, and the vector field $\mathbf{A} = (A_1, A_2)$ is given and satisfies

$$\text{curl } \mathbf{A} \equiv \partial_1 A_2 - \partial_2 A_1 = 1 \quad \text{and} \quad \text{div } \mathbf{A} = 0 \quad \text{in } \Omega, \qquad \mathbf{A}\cdot\nu = 0 \quad \text{on } \partial\Omega. \tag{3.2}$$

From the variational characterization, the lowest eigenvalue is given by

$$\beta_\gamma(\varepsilon^{-2}\mathbf{A}) = \inf_{\phi\in W^{1,2}(\Omega)} \frac{\int_\Omega |\nabla_{\varepsilon^{-2}\mathbf{A}}\phi|^2 dx + \gamma\int_{\partial\Omega}|\phi|^2 ds}{\int_\Omega |\phi|^2 dx}.$$

THEOREM 3.1. *Let* $\mathbf{A}$ *satisfy* (3.2) *and let* $\gamma > 0$ *be fixed. We have, as* $\varepsilon \to 0$,

$$\beta_\gamma(\varepsilon^{-2}\mathbf{A}) = \frac{1}{\varepsilon^2}[\beta_0 + C_1(3\gamma - \kappa_{\max})\varepsilon + O(\varepsilon^{4/3})], \tag{3.3}$$

where β_0 *and* C_1 *are the numbers given in section* 2, *and* $\kappa_{\max}$ *is the maximum value of the curvature of* $\partial\Omega$.

Proof. An upper bound of H_{C_3} with different error terms, which can be controlled by careful computations, has been obtained by [LP4, Lemma A.3].[10] When $\gamma = 0$, the lower bound estimate was obtained by Helffer and Morame [HM1]. Note that the key ingredients in Helffer and Morame's arguments are (i) the estimates of various cut-off functions; (ii) the estimate of the lowest eigenvalue of an ordinary differential operator with parameter ε in the half-line $\mathbb{R}_+$, with a homogeneous Neumann boundary condition $u'(0) = 0$. When $\gamma > 0$, (i) is still valid, and (ii) needs a minor modification to fit the de Gennes boundary condition $u'(0) + \varepsilon\gamma u(0) = 0$, and hence we can modify Helffer and Morame's arguments to get the lower bound. The details are omitted. □

[10] The constant C_2 appeared in [LP4] and was given by $\frac{u(0)^2}{\|u\|^2_{L^2}}$, where u is the eigenfunction of (2.1) for $z = z_0$. Hence $C_2 = 3C_1$.

As in [HM1], we can show that the eigenfunctions of (3.1) concentrate at the set of the maximum points of the curvature of $\partial\Omega$, $\mathcal{N}(\partial\Omega)$, as $\varepsilon \to 0$. In a neighborhood of $\partial\Omega$, $\Omega_\delta = \{x : \text{dist}(x, \partial\Omega) < \delta\}$, where $\delta > 0$ is small, $d_{\partial\Omega}(x)$ is a smooth function. Thus we can introduce a local coordinate (s, t), with s measuring the tangential distance and $t = d_{\partial\Omega}(x)$ measuring the normal distance. Any point $x \in \Omega_\delta$ can be represented by a vector-valued function $x(s, t)$, with $|s| \le L = |\partial\Omega|/2$ and $0 \le t < \delta$. Thus a point $x(s, t) \in \Omega_\delta$ corresponds with a unique (s, t), and hence we have functions $s(x)$ and $t(x)$, which are well defined on Ω_δ. After extending these functions onto $\bar\Omega$, every point $x \in \bar\Omega$ can be represented as $x = x(s, t)$.

COROLLARY 3.2. *Let* $\mathbf{A}$ *satisfy* (3.2) *and let* $\gamma > 0$ *be fixed. There exist positive constants* α_j, c_j, *and* M_j *depending only on* Ω *such that for the eigenfunctions* ϕ_ε *of* (3.1) *associated with* $\mu_\gamma(\varepsilon^{-2}\mathbf{A})$, *we have*

(i) $\int_\Omega \exp\big(\alpha_1\varepsilon^{-1}d_{\partial\Omega}(x)\big)|\phi_\varepsilon|^2dx \le M_1 \int_{\Omega\cap\{d_{\partial\Omega}(x)<c_2\varepsilon\}} |\phi_\varepsilon|^2dx$;

(ii) $\int_\Omega \exp\big(\alpha_2\varepsilon^{-1/2}[\kappa_{\max} - \kappa_r(s) - c_1\varepsilon^{1/3}]\big)|\phi_\varepsilon|^2dx \le M_2 \int_{\Omega\cap\{d_{\partial\Omega}(x)<c_2\varepsilon\}} |\phi_\varepsilon|^2dx$.

Proof. Conclusion (i) has been established in [HM1] when $\gamma = 0$, and the proof works when $\gamma > 0$, with a slight modification. In order to prove (ii), we use the method in [HM1] and (3.3) to show that for any $\mathbf{A}$ satisfying (3.2) and $\gamma > 0$, there exists a positive constant C such that for any $\phi \in W^{1,2}(\Omega)$ and small $\varepsilon > 0$,

$$\int_\Omega |\nabla_{\varepsilon^{-2}\mathbf{A}}\phi|^2dx + \gamma\int_{\partial\Omega} |\phi|^2ds \ge \frac{1}{\varepsilon^2}\int_\Omega W_\Omega(x)|\phi|^2dx, \tag{3.4}$$

where

$$W_\Omega(x) = \begin{cases} 1 - C\varepsilon^{4/3} & \text{if } d_{\partial\Omega}(x) \ge 2\varepsilon^{1/3}, \\ \beta_0 + C_1[3\gamma - \kappa_r(s)]\varepsilon - C\varepsilon^{4/3} & \text{if } d_{\partial\Omega}(x) < 2\varepsilon^{1/3}. \end{cases} \tag{3.5}$$

Here β_0 and C_1 are the numbers given in section 2, and $s = s(x)$ is associated with x through the representation $x = x(s, t)$. Then we apply the argument in [HP, proof of Theorem 6.1] and use (3.4) to get (ii). □

4. Elliptic estimates of minimizers. The regularity of the weak solutions of the Ginzburg–Landau system in three-dimensional domains has been discussed in [P4]. In this section we establish the estimates on a film, with constants independent of the thickness of the film. Let

$$D^{1,2}(\mathbb{R}^3) = \{\phi \in L^2_{\text{loc}}(\mathbb{R}^3) : \ \nabla\phi \in L^2(\mathbb{R}^3)\},$$

and on $D^{1,2}(\mathbb{R}^3)$ we define a seminorm $\|\phi\|_{1,2} = \|\nabla\phi\|_{L^2(\mathbb{R}^3)}$. After identifying two functions that differ by a constant, $(D^{1,2}(\mathbb{R}^3), \|\cdot\|_{1,2})$ is a Hilbert space [G, Lemma II.5.1]. Let $\mathbf{D}^{1,2}(\mathbb{R}^3)$ denote the corresponding space of vector fields. It follows from Theorem II.6.2 in [G] that for any $\mathbf{B} \in \mathbf{D}^{1,2}(\mathbb{R}^3)$, there exists a unique constant vector $\mathbf{b}$ such that $\mathbf{B} - \mathbf{b}$ can be approximated in the norm $\|\cdot\|_{1,2}$ by C_0^∞ vector fields. It is well known that for any $\mathbf{B} \in \mathbf{D}^{1,2}(\mathbb{R}^3)$,

$$\|\mathbf{B}\|^2_{1,2} \equiv \int_{\mathbb{R}^3} |\nabla\mathbf{B}|^2dxdz = \int_{\mathbb{R}^3} \{|\text{curl } \mathbf{B}|^2 + |\text{div } \mathbf{B}|^2\}dxdz; \tag{4.1}$$

see [L], as well as [GP, (3.2)]. The following space will also be useful to us:

$$\mathbf{D}^{1,2}(\mathbb{R}^3, \text{div}) = \{\mathbf{B} \in \mathbf{D}^{1,2}(\mathbb{R}^3) : \ \text{div } \mathbf{B} = 0 \ \text{ in } \mathbb{R}^3\}.$$

Let $W^{1,2}(D_l,\mathbb{C})$ be the Sobolev space of all complex-valued functions defined on $\bar{D}_l$. Given a unit vector $\mathbf{h}$, define $\mathbf{F_h}$ by (1.2), and let

$$\mathcal{W}(D_l)=\{(\psi,\mathbf{A}):\ \psi\in W^{1,2}(D_l,\mathbb{C}),\ \mathbf{A}-\mathbf{F_h}\in\mathbf{D}^{1,2}(\mathbb{R}^3)\}. \tag{4.2}$$

We consider the variational problem for the Ginzburg–Landau functional $\mathcal{G}$ on $\mathcal{W}(D_l)$. Define

$$C(\mathbf{h},\kappa,l,\sigma)=\inf_{(\psi,\mathbf{A})\in\mathcal{W}(D_l)}\mathcal{G}[\psi,\mathbf{A}]. \tag{4.3}$$

Due to the gauge invariance of $\mathcal{G}$, we can replace $\mathbf{A}$ by $\hat{\mathbf{A}}$ that satisfies curl $\hat{\mathbf{A}}=$ curl $\mathbf{A}$ and div $\hat{\mathbf{A}}=0$ in $\mathbb{R}^3$; see [GP, Lemma 3.1]. So we can restrict the functional $\mathcal{G}$ on a subspace of $\mathcal{W}(D_l)$:

$$\mathcal{W}(D_l,\mathrm{div})=\{(\psi,\mathbf{A})\in\mathcal{W}(D_l):\ \mathrm{div}\ \mathbf{A}=0\ \text{in}\ \mathbb{R}^3\}. \tag{4.4}$$

It is easy to show that the (global) minimizers exist, and they are the weak solutions of the Euler equations

$$\begin{cases} -\nabla^2_{\kappa\sigma\mathbf{A}}\psi=\kappa^2(1-|\psi|^2)\psi & \text{in } D_l,\\ \mathrm{curl}^2\mathbf{A}=\dfrac{1}{\kappa\sigma}\Im\{\bar{\psi}\nabla_{\kappa\sigma\mathbf{A}}\psi\}\chi_{D_l} & \text{in } \mathbb{R}^3,\\ (\nabla_{\kappa\sigma\mathbf{A}}\psi)\cdot\nu_D+\gamma\psi=0 & \text{on } \partial^*D_l,\\ \mathbf{A}-\mathbf{F_h}\in\mathbf{D}^{1,2}(\mathbb{R}^3,\mathrm{div}), \end{cases} \tag{4.5}$$

where χ_{D_l} is the characteristic function of D_l, namely, χ_{D_l} equals 1 in D_l and equals 0 in $\mathbb{R}^3\setminus D_l$. In the content of weak solutions, (4.5) is equivalent to (1.1); see [L, Chapter 5, section 4].

LEMMA 4.1. *Let $\kappa>0$, $\sigma>0$, and let $(\psi,\mathbf{A})$ be a minimal solution of* (1.1). *Then, for any $0<\alpha<1$,*

$$\begin{aligned} &\psi\in C^{2+\alpha}(D_l\cup\partial\Omega\times[0,l])\bigcup C^{2+\alpha}(D_l\cup\Omega\times\{0,l\}),\\ &\mathbf{A}\in C^{1+\alpha}(\bar{D}_l)\cup C^{2+\alpha}_{loc}(D_l)\cup C^{2+\alpha}_{loc}(\mathbb{R}^3\setminus\bar{D}_l). \end{aligned} \tag{4.6}$$

Moreover, we have the following estimates:

(i) *There exists $C>0$ independent of $\mathbf{h}$, l, κ, and σ such that*

$$\begin{aligned} \|\mathbf{A}-\mathbf{F_h}\|_{L^6(\mathbb{R}^3)}&\le\frac{C}{\sigma}\|\psi\|^2_{L^4(D_l)},\\ \|\mathrm{curl}\ \mathbf{A}-\mathbf{h}\|_{L^2(\mathbb{R}^3)}&\le\frac{C}{\sigma}\|\psi\|^2_{L^4(D_l)}. \end{aligned} \tag{4.7}$$

(ii) *For any $0<\alpha<1$, and for any $R>0$ such that $\bar{D}_l\subset B_R$, there exists $C(\alpha,R)>0$ independent of $\mathbf{h}$, l, κ, and σ such that for $q=3/(1-\alpha)$,*

$$\|\mathbf{A}-\mathbf{F_h}\|_{C^{1+\alpha}(R_R)}\le\ C(\alpha,R)\left\{\frac{1}{\sigma}\|\psi\|^2_{L^4(D_l)}+\frac{1}{\kappa\sigma}\|\Im\{\bar{\psi}\nabla_{\kappa\sigma\mathbf{A}}\psi\}\|_{L^q(D_l)}\right\}. \tag{4.8}$$

(iii) *For any positive numbers $a < b$ and $q > 1$, there exist positive constants $C(a,b)$ and $C(a,b,q)$ independent of $\mathbf{h}$, l, κ, and σ ($a\kappa^{-1} \le \sigma \le b\kappa$) such that*

$$\begin{aligned} &\|\Im\{\bar\psi\nabla_{\kappa\sigma\mathbf{A}}\psi\}\|_{L^\infty(D_l)} \le C(a,b)\sqrt{\kappa\sigma}\|\psi\|^2_{L^\infty(D_l)},\\ &\|\Im\{\bar\psi\nabla_{\kappa\sigma\mathbf{A}}\psi\}\|_{L^q(D_l)} \le C(a,b,q)\sqrt{\kappa\sigma}\|\psi\|_{L^q(D_l)}\|\psi\|_{L^\infty(D_l)}. \end{aligned} \tag{4.9}$$

Proof. Step 1. We prove (4.7). From the first equation in (1.1) we have

$$\int_{D_l} |\nabla_{\kappa\sigma\mathbf{A}}\psi|^2 dxdz = \kappa^2 \int_{D_l} (1-|\psi|^2)|\psi|^2 dxdz.$$

So

$$\mathcal{G}[\psi,\mathbf{A}] = \frac{\kappa^2}{2}|D_l| - \frac{\kappa^2}{2}\int_{D_l}|\psi|^4 dxdz + \kappa^2\sigma^2\int_{\mathbb{R}^3}|\text{curl }\mathbf{A}-\mathbf{h}|^2 dxdz.$$

Since $C(\mathbf{h},\kappa,l,\sigma) \le \mathcal{G}[0,\mathbf{F_h}] = \frac{\kappa^2}{2}|D_l|$, we have

$$\kappa^2\sigma^2\int_{\mathbb{R}^3}|\text{curl }\mathbf{A}-\mathbf{h}|^2 dxdz = C(\mathbf{h},\kappa,l,\sigma) - \frac{\kappa^2}{2}|D_l| + \frac{\kappa^2}{2}\int_{D_l}|\psi|^4 dxdz \le \frac{\kappa^2}{2}\|\psi\|^4_{L^4(\Omega)}.$$

So the second inequality in (4.7) is true. Using the Sobolev imbedding theorem we have

$$\|\mathbf{A}-\mathbf{F_h}\|^2_{L^6(\mathbb{R}^3)} \le C\int_{\mathbb{R}^3}\{|\text{curl}(\mathbf{A}-\mathbf{F_h})|^2 + |\text{div}(\mathbf{A}-\mathbf{F_h})|^2\}dxdz.$$

Since $\text{div}(\mathbf{A}-\mathbf{F_h}) = 0$, the first inequality in (4.7) follows.

Step 2. We prove (4.8). Let

$$\mathbf{f} = \frac{1}{\kappa\sigma}\Im\{\bar\psi\nabla_{\kappa\sigma\mathbf{A}}\psi\}\chi_{D_l}, \qquad \mathbf{U} = \mathbf{A}-\mathbf{F_h}.$$

$\mathbf{f} \in L^2(\mathbb{R}^3)$, and $\mathbf{U}$ is a weak solution of the equation $\text{curl}^2\mathbf{U} = \mathbf{f}$ in $\mathbb{R}^3$. Since $\text{div }\mathbf{U} = 0$, this equation can be written as

$$-\Delta\mathbf{U} = \mathbf{f} \qquad \text{in } \mathbb{R}^3. \tag{4.10}$$

Applying the De Giorge L^∞ estimate [GT, Theorem 8.17] to each component of (4.10), we find that there exists $C > 0$ independent of κ and σ such that for any $0 < r < R$,

$$\|\mathbf{U}\|_{L^\infty(B_r)} \le C\big\{(R-r)^{-3/2}\|\mathbf{U}\|_{L^2(B_R)} + R^{1/2}\|\mathbf{f}\|_{L^2(B_R)}\big\}.$$

Let us choose $\rho > \max_{(x,z)\in D_l}\sqrt{|x|^2+z^2}$ and $\rho < r < R$. We have

$$\|\mathbf{U}\|_{L^\infty(B_r)} \le C(r,R)\big\{\|\mathbf{U}\|_{L^2(B_R)} + \|\mathbf{f}\|_{L^2(B_R)}\big\}.$$

Then, applying the Hölder estimate for weak solutions [GT, Theorem 8.24], we find that for some $0 < \alpha_0 < 1$ and for all $\rho < r < R$,

$$\|\mathbf{U}\|_{C^{\alpha_0}(B_r)} \le C(r,R)\big\{\|\mathbf{U}\|_{L^2(B_R)} + \|\mathbf{f}\|_{L^2(B_R)}\big\}. \tag{4.11}$$

In particular, $\mathbf{A} \in C^{\alpha_0}(\bar D_l)$.

Next, from (1.1) we see that ψ satisfies

$$\begin{cases} -\Delta\psi + 2i\kappa\sigma\mathbf{A}\cdot\nabla\psi + \kappa^2\sigma^2|\mathbf{A}|^2\psi = \kappa^2(1-|\psi|^2)\psi & \text{in } D_l, \\ \dfrac{\partial\psi}{\partial\nu_D} - i\kappa\sigma(\mathbf{A}\cdot\nu_D)\psi + \gamma\psi = 0 & \text{on } \partial^* D_l. \end{cases} \tag{4.12}$$

Since $\mathbf{A} \in C^{\alpha_0}(\bar{D}_l)$, applying the Hölder estimate (see, for instance, [GT, Lemma 6.27]) on $\bar{D}_l$, we find that

$$\psi \in C^{1+\alpha_0}(D_l \cup \partial\Omega \times [0,l]) \bigcup C^{1+\alpha_0}(D_l \cup \Omega \times \{0,l\}).$$

Therefore,

$$\mathbf{f} \in C^{\alpha_0}(D_l \cup \partial\Omega \times [0,l]) \bigcup C^{\alpha_0}(D_l \cup \Omega \times \{0,l\}),$$

and hence $\mathbf{f} \in L^\infty(\mathbb{R}^3)$.

Now we apply the Hölder gradients estimate to (4.10) again and find that for any $0 < \alpha < 1$, $\mathbf{U} \in C^{1+\alpha}_{\text{loc}}(\mathbb{R}^3)$ (hence $\mathbf{A} \in C^{1+\alpha}_{\text{loc}}(\mathbb{R}^3)$), and for $q = 3/(1-\alpha)$,

$$\|\nabla\mathbf{U}\|_{C^\alpha(B_r)} \le C(r,R,\alpha)\big\{\|\mathbf{f}\|_{L^q(B_R)} + \|\nabla\mathbf{U}\|_{L^2(B_R)} + \|\mathbf{U}\|_{L^2(B_R)}\big\}.$$

Since div $\mathbf{A} = 0$, we find that

$$\|\nabla\mathbf{U}\|_{C^\alpha(B_r)} \le C(r,R,\alpha)\big\{\|\mathbf{f}\|_{L^q(B_R)} + \|\text{curl } \mathbf{U}\|_{L^2(B_R)} + \|\mathbf{U}\|_{L^2(B_R)}\big\}.$$

Using this inequality and $\|\mathbf{U}\|_{L^2(B_R)} \le CR\|\mathbf{U}\|_{L^6(B_R)}$, together with (4.11) and (4.7), we get (4.8).

Step 3. For any $0 < \alpha < 1$, since $\mathbf{A} \in C^{1+\alpha}(\bar{D}_l)$, we apply the Hölder estimate to (4.12) again (see [GT, Lemma 6.27]) and find that

$$\psi \in C^{2+\alpha}(D_l \cup \Omega \times [0,l]) \bigcup C^{2+\alpha}(D_l \cup \partial\Omega \times \{0,l\}).$$

Step 4. We prove (4.9). First we assume $l \ge (\kappa\sigma)^{-1/2}$. Without loss of generality we assume $\kappa\sigma \gg 1$. We use a blow-up argument in the scale of $(\kappa\sigma)^{-1/2}$ and get a limiting equation with bounded coefficients and apply the L^∞ estimate to it. Then we return to the original scale and find

$$\|\nabla_{\kappa\sigma\mathbf{A}}\psi\|_{L^\infty(D_l)} \le C\sqrt{\kappa\sigma}\|\psi\|_{L^\infty(D_l)}. \tag{4.13}$$

Equation (4.9) follows from (4.13) immediately. We omit the details but refer to [HP, Proposition 4.2] for a two-dimensional problem.

Next we assume that $l \le (\kappa\sigma)^{-1/2}$. Recall that div $\mathbf{A} = 0$. Let $\chi \in W^{1,2}(D_l)$ be a weak solution of the equation

$$\Delta\chi = 0 \quad \text{in } D_l, \qquad \frac{\partial\chi}{\partial\nu_D} = \mathbf{A}\cdot\nu_D \quad \text{on } \partial^* D_l,$$

and satisfy $\int_{D_l}\chi dxdz = 0$. Note that χ is a piecewise $C^{2+\alpha}$ function, and (see [GT, Lemma 6.27])

$$\begin{aligned} &\|\chi\|_{C^{2+\alpha}(D_l \cup \partial\Omega\times[0,l])} + \|\chi\|_{C^{2+\alpha}(D_l \cup \bar{\Omega}\times\{0,l\})} \\ &\le C\Big\{\|\mathbf{A}\cdot\nu_D\|_{C^{1+\alpha}(D_l \cup \partial\Omega\times[0,l])} + \|\mathbf{A}\cdot\nu_D\|_{C^{1+\alpha}(D_l \cup \bar{\Omega}\times\{0,l\})}\Big\} \le C\|\mathbf{A}\|_{C^{1+\alpha}(B_\rho)}, \end{aligned} \tag{4.14}$$

where $\rho > \max_{(x,z)\in D_l} \sqrt{|x|^2+z^2}$.

Let $\tilde{\psi} = e^{-i\kappa\sigma\chi}\psi$ and $\tilde{\mathbf{A}} = \mathbf{A} - \nabla\chi$. Then $\tilde{\mathbf{A}}\cdot\nu_D = 0$ on $\partial^* D_l$. In particular,

$$\tilde{A}_3(x_1,x_2,0) = \tilde{A}_3(x_1,x_2,l) = 0. \tag{4.15}$$

$\tilde{\psi}$ satisfies

$$\begin{cases} -\Delta\tilde{\psi} + 2i\kappa\sigma\tilde{\mathbf{A}}\cdot\nabla\tilde{\psi} + \kappa^2\sigma^2|\tilde{\mathbf{A}}|^2\tilde{\psi} = \kappa^2(1-|\tilde{\psi}|^2)\tilde{\psi} & \text{in } D_l, \\ \dfrac{\partial\tilde{\psi}}{\partial\nu_D} + \gamma\tilde{\psi} = 0 & \text{on } \partial^* D_l. \end{cases}$$

Let ξ_l be the eigenfunction of (2.3) and $c_l = 1/\|\xi_l\|_{L^\infty([0,l])}$. Using (2.4) and (2.5) we can find $C>0$ such that for all small l,

$$\max_{0\le z\le l}\left|\frac{\xi_l'(z)}{\xi_l(z)}\right| \le C.$$

Let

$$\phi = \frac{\tilde{\psi}}{c_l\xi_l} = \frac{e^{-i\chi}}{c_l\xi_l}\psi.$$

Then ϕ satisfies

$$\begin{cases} -\Delta\phi + 2i\kappa\sigma\tilde{\mathbf{A}}\cdot\nabla\phi + \kappa^2\sigma^2|\tilde{\mathbf{A}}|^2\phi - \dfrac{2\xi_l'}{\xi_l}\partial_z\phi + \Big[\tau_\gamma(l)^2 + 2i\kappa\sigma\tilde{A}_3\dfrac{\xi_l'}{\xi_l}\Big]\phi \\ \qquad = \kappa^2(1-c_l^2\xi_l^2|\phi|^2)\phi & \text{in } D_l, \\ \dfrac{\partial\phi}{\partial z} = 0 & \text{if } z=0, \text{ or } l, \\ \dfrac{\partial\phi}{\partial\nu} + \gamma\phi = 0 & \text{on } \partial\Omega\ \times(0,l). \end{cases} \tag{4.16}$$

Using (4.15), we can extend ϕ and $\tilde{\mathbf{A}}$ in the z direction in the following way: for $0<z<l$,

$$\begin{aligned} &\tilde{A}_j(x_1,x_2,-z) = \tilde{A}_j(x_1,x_2,z) \quad \text{and} \quad \tilde{A}_j(x_1,x_2,l+z) = \tilde{A}_j(x_1,x_2,l-z), \quad j=1,\ 2, \\ &\tilde{A}_3(x_1,x_2,-z) = \tilde{A}_3(x_1,x_2,z) \quad \text{and} \quad \tilde{A}_3(x_1,x_2,l+z) = -\tilde{A}_3(x_1,x_2,l-z), \\ &\phi(x_1,x_2,-z) = \phi(x_1,x_2,z) \qquad \text{and} \quad \phi(x_1,x_2,l+z) = \phi(x_1,x_2,l-z). \end{aligned}$$

Extend ξ_l by letting $\xi_l(-z) = \xi_l(z)$ and $\xi_l(l+z) = \xi_l(l-z)$. Then we get an equation for ϕ on $\Omega\times(-l,2l)$, which is an extension of (4.16):

$$\begin{cases} -\Delta\phi + 2i\kappa\sigma\tilde{\mathbf{A}}\cdot\nabla\phi + \kappa^2\sigma^2|\tilde{\mathbf{A}}|^2\phi - \dfrac{2\xi_l'}{\xi_l}\partial_z\phi + \Big[\tau_\gamma(l)^2 + 2i\kappa\sigma\tilde{A}_3\dfrac{\xi_l'}{\xi_l}\Big]\phi \\ \qquad = \kappa^2(1-c_l^2\xi_l^2|\phi|^2)\phi & \text{in } \Omega\times(-l,2l), \\ \dfrac{\partial\phi}{\partial z} = 0 & \text{if } z=-l, \text{ or } 2l, \\ \dfrac{\partial\phi}{\partial\nu} + \gamma\phi = 0 & \text{on } \partial\Omega\ \times(-l,2l). \end{cases} \tag{4.17}$$

We can verify that $\tilde{\mathbf{A}} \in C^\alpha(\bar{\Omega}\times[-l,2l])$. After the extension ξ_l' is no longer continuous at $z=0$ and l, however, $\xi_l'\partial_z\phi$ and $\tilde{A}_3\xi_l'$ vanish at $z=0$ and l. So we can verify that ϕ is a weak solution of (4.17) on $\Omega\times(-l,2l)$.

We can repeat this process and extend ϕ onto a cylinder $\Omega\times(a,b)$ with $b-a=1$, and ϕ is a weak solution of a new equation similar to (4.17) on the cylinder. Then we apply the elliptic estimate to the new equation and conclude that

$$\|\nabla_{\kappa\sigma\tilde{\mathbf{A}}}\phi\|_{L^\infty(\bar{\Omega}\times[a,b])} \le C\sqrt{\kappa\sigma}\|\phi\|_{L^\infty(\Omega\times[a,b])} = C\sqrt{\kappa\sigma}\|\psi\|_{L^\infty(D_l)}.$$

Returning to D_l we get

$$\|\nabla_{\kappa\sigma\tilde{\mathbf{A}}}\phi\|_{L^\infty(D_l)} \le C\sqrt{\kappa\sigma}\|\psi\|_{L^\infty(D_l)}.$$

Since $\psi = c_l\xi_l\phi e^{i\kappa\sigma\chi}$ and $\mathbf{A} = \tilde{\mathbf{A}} + \nabla\chi$, we have

$$\begin{aligned}
&\bar{\psi}\nabla_{\kappa\sigma\mathbf{A}}\psi = c_l^2\xi_l^2\bar{\phi}\nabla_{\kappa\sigma\tilde{\mathbf{A}}}\phi + c_l^2|\phi|^2\xi_l\nabla\xi_l,\\
&\Im\{\bar{\psi}\nabla_{\kappa\sigma\mathbf{A}}\psi\} = c_l^2\xi_l^2\Im\{\bar{\phi}\nabla_{\kappa\sigma\tilde{\mathbf{A}}}\phi\}.
\end{aligned}$$

Hence

$$\begin{aligned}
&\|\Im\{\bar{\psi}\nabla_{\kappa\sigma\mathbf{A}}\psi\}\|_{L^\infty(D_l)} \le \|\Im\{\bar{\phi}\nabla_{\kappa\sigma\tilde{\mathbf{A}}}\phi\}\|_{L^\infty(D_l)} \le C\sqrt{\kappa\sigma}\|\psi\|^2_{L^\infty(D_l)},\\
&\|\Im\{\bar{\psi}\nabla_{\kappa\sigma\mathbf{A}}\psi\}\|_{L^q(D_l)} \le \|\Im\{\bar{\phi}\nabla_{\kappa\sigma\tilde{\mathbf{A}}}\phi\}\|_{L^q(D_l)} \le C(q)\sqrt{\kappa\sigma}\|\psi\|_{L^q(D_l)}\|\psi\|_{L^\infty(D_l)}.
\end{aligned}$$

Equation (4.9) is proved. □

5. An eigenvalue problem on films. In this section we study the lowest eigenvalue $\mu_\gamma(\varepsilon^{-2}\mathbf{F})$ of

$$\begin{cases} -\nabla^2_{\varepsilon^{-2}\mathbf{F}}\phi = \mu\phi & \text{in } D_l,\\ (\nabla_{\varepsilon^{-2}\mathbf{F}}\phi)\cdot\nu_D + \gamma\phi = 0 & \text{on } \partial^* D_l, \end{cases} \tag{5.1}$$

where curl $\mathbf{F} = \mathbf{e}_3 = (0,0,1)$.

THEOREM 5.1. *Let* curl $\mathbf{F} = (0,0,1)$. *For small* ε *and* l *we have*

$$\mu_\gamma(\varepsilon^{-2}\mathbf{F}) = \tau_\gamma(l)^2 + \beta_0\varepsilon^{-2} + C_1(3\gamma - \kappa_{\max})\varepsilon^{-1} + O(\varepsilon^{-2/3}), \tag{5.2}$$

where $\tau_\gamma(l)^2$ *is the lowest eigenvalue of* (2.3), *and* β_0 *and* C_1 *are the numbers given in section* 2.

Proof. Due to gauge invariance of the operator $-\nabla^2_{\varepsilon^{-2}\mathbf{F}}$, we choose $\mathbf{F} = (-x_2,0,0)$. Then (5.1) is a separable equation. The eigenfunctions have the form $\phi(x,z) = \varphi(x_1,x_2)\xi_l(z)$, and

$$\mu_\gamma(\varepsilon^{-2}\mathbf{F}) = \beta_\gamma(\varepsilon^{-2}\mathbf{E}) + \tau_\gamma(l)^2,$$

where ξ_l is the function given in (2.4), $\mathbf{E} = (-x_2,0)$, and $\beta_\gamma(\varepsilon^{-2}\mathbf{E})$ is the lowest eigenvalue of (3.1) for $\mathbf{A} = \mathbf{E}$. An estimate for $\beta_\gamma(\varepsilon^{-2}\mathbf{E})$ has been obtained in (3.3). So the conclusion follows. □

Next we establish an integral inequality for functions vanishing at the lateral surface of D_l. Let

$$W^{1,2}(D_l,0) = \{\phi\in W^{1,2}(D_l):\ \phi = 0 \ \text{ on } \partial\Omega\times[0,l]\}.$$

LEMMA 5.2. *Let* $\mathbf{A} = (A_1, A_2, A_3) \in C^1(\bar{D}_l)$. *For any* $\phi \in W^{1,2}(D_l, 0)$ *we have*

(5.3)
$$\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}}\phi|^2 dxdz + \gamma \int_{\partial D_l} |\phi|^2 dS \geq \int_{D_l} \{\varepsilon^{-2}(\partial_1 A_2 - \partial_2 A_1) + \tau_\gamma(l)^2\}|\phi|^2 dxdz.$$

In particular, for any $\phi \in W^{1,2}(D_l, 0)$,

$$\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{F}}\phi|^2 dxdz + \gamma \int_{D_l} |\phi|^2 dS \geq \left[\varepsilon^{-2} + \tau_\gamma(l)^2\right] \int_{D_l} |\phi|^2 dxdz. \tag{5.4}$$

Proof. Let $\phi \in W^{1,2}(D_l, 0) \cap C^1(\bar{D}_l)$. For a fixed $z \in [0, l]$, since $\phi(x, z) = 0$ for $x \in \partial\Omega$, using Theorem 4 in [M] we have

$$\int_\Omega \left\{|(\partial_1 - i\varepsilon^{-2}A_1)\phi(x,z)|^2 + |(\partial_2 - i\varepsilon^{-2}A_2)\phi(x,z)|^2\right\} dx \geq \frac{1}{\varepsilon^2} \int_\Omega (\partial_1 A_2 - \partial_2 A_1)|\phi(x,z)|^2 dx.$$

Integrating this inequality in z we get

$$\int_{D_l} \left\{|(\partial_1 - i\varepsilon^{-2}A_1)\phi|^2 + |(\partial_2 - i\varepsilon^{-2}A_2)\phi|^2\right\} dxdz \geq \frac{1}{\varepsilon^2} \int_{D_l} (\partial_1 A_2 - \partial_2 A_1)|\phi|^2 dxdz.$$

On the other hand, for a fixed $x \in \Omega$, using the Kato's inequality we find

$$\begin{aligned} &\int_0^l |(\partial_z - i\varepsilon^{-2}A_3)\phi(x,z)|^2 dz + \gamma\left[|\phi(x,0)|^2 + |\phi(x,l)|^2\right] \\ &\geq \int_0^l \left|\partial_z|\phi(x,z)|\right|^2 dz + \gamma\left[|\phi(x,0)|^2 + |\phi(x,l)|^2\right] \geq \tau_\gamma(l)^2 \int_0^l |\phi(x,z)|^2 dz. \end{aligned}$$

Thus

$$\int_{D_l} |(\partial_z - i\varepsilon^{-2}A_3)\phi|^2 dxdz + \gamma \int_\Omega \left[|\phi(x,0)|^2 + |\phi(x,l)|^2\right] dx \geq \tau_\gamma(l)^2 \int_{D_l} |\phi|^2 dxdz.$$

So we get (5.3). Since $\partial_1 F_2 - \partial_2 F_1 = 1$, (5.4) follows. □

The inequality (5.4) yields an estimate of the lowest eigenvalue of $-\nabla^2_{\varepsilon^{-2}\mathbf{F}}$ in D_l with Dirichlet boundary condition on the lateral surface $\partial\Omega \times [0, l]$ and Robin condition on the top and bottom faces:

$$\begin{cases} -\Delta\phi + 2i\varepsilon^{-2}x_2\partial_1\phi + \varepsilon^{-4}x_2^2\phi = \lambda\phi & \text{in } \Omega \times (0, l), \\ \partial_z\phi = \gamma\phi & \text{on } \Omega \times \{0\}, \\ \partial_z\phi = -\gamma\phi & \text{on } \Omega \times \{l\}, \\ \phi = 0 & \text{on } \partial\Omega \times (0, l). \end{cases}$$

This problem can be solved by the method of separable variables as for (5.1). Hence the lowest eigenvalue is given by $\lambda = \alpha + \tau_\gamma(l)^2$, where α is the lowest eigenvalue of the Dirichlet problem

$$\begin{cases} -\Delta\varphi + 2i\varepsilon^{-2}x_2\partial_1\varphi + \varepsilon^{-4}x_2^2\varphi = \alpha\varphi & \text{in } \Omega, \\ \varphi = 0 & \text{on } \partial\Omega. \end{cases}$$

Using the result in [LP2, Theorem 2] about an eigenvalue problem in the entire plane we find that $\alpha \geq \varepsilon^{-2}$. So $\lambda \geq \varepsilon^{-2} + \tau_\gamma(l)^2$. Hence we get (5.4) again.

Now we can establish an integral inequality on D_l, which can be viewed as a three-dimensional version of (3.4) and is useful in the study of the concentration phenomenon of order parameters on films.

THEOREM 5.3. *For any $\phi \in W^{1,2}(D_l)$ we have*

$$\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{F}}\phi|^2 dxdz + \gamma \int_{\partial D_l} |\phi|^2 dS \geq \varepsilon^{-2} \int_{D_l} \mathcal{W}_{D_l}(x,z)|\phi|^2 dxdz. \tag{5.5}$$

$\mathcal{W}_{D_l}$ is defined by

$$\mathcal{W}_{D_l}(x,z) = \begin{cases} 1 + \varepsilon^2\tau_\gamma(l)^2 - C\varepsilon^{4/3} & \text{if } d_{\partial\Omega}(x) \geq 2\varepsilon^{1/3}, \\ \beta_0 + C_1[3\gamma - \kappa_r(s)]\varepsilon + \varepsilon^2\tau_\gamma(l)^2 - C\varepsilon^{4/3} & \text{if } d_{\partial\Omega}(x) < 2\varepsilon^{1/3}, \end{cases} \tag{5.6}$$

where β_0, C_1, and $\tau_\gamma(l)$ are the numbers given in section 2, C depends only on Ω and γ, and $s = s(x)$ is associated with x through the representation $x = x(s,t)$.

Proof. We apply the idea in [HM1]. Let (s,t) be the local coordinates in Ω_δ described in section 3, and (s,t) has been extended onto $\bar{\Omega}$. Then (s,t,z) gives the new coordinates on $\bar{D}_l$. Choose cut-off functions $\eta_0(t)$ and $\eta_1(t)$ depending only on t such that

$$\mathrm{spt}(\eta_0) \subset \left[\frac{\varepsilon^{1/3}}{20}, +\infty\right), \quad \mathrm{spt}(\eta_1) \subset \left(-\infty, \frac{\varepsilon^{1/3}}{10}\right], \quad |\eta_j'(t)| \leq C\varepsilon^{-1/3}, \quad \eta_0^2 + \eta_1^2 = 1.$$

Then

$$\begin{aligned} &\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{F}}\phi|^2 dxdz + \gamma \int_{\partial D_l} |\phi|^2 dS \\ &= \sum_{j=0}^{1} \left\{ \int_{D_l} \left[|\nabla_{\varepsilon^{-2}\mathbf{F}}(\eta_j\phi)|^2 - |\nabla\eta_j|^2|\phi|^2 \right] dxdz + \gamma \int_{\partial D_l} |\eta_j\phi|^2 dS \right\}. \end{aligned} \tag{5.7}$$

Note that $\eta_0\phi$ is supported in $\Omega^\varepsilon \times [0,l]$, where

$$\Omega^\varepsilon = \left\{ x \in \Omega : \ d_{\partial\Omega}(x) \geq \frac{\varepsilon^{1/3}}{20} \right\},$$

namely, $\eta_0\phi = 0$ in a thin cylinder around the lateral surface $\partial\Omega \times [0,l]$. $\eta_1\phi$ is supported near the lateral surface.

Step 1. We estimate $\eta_0\phi$. The sum in (5.7) involving $\eta_0\phi$ is

$$S_1 \equiv \int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{F}}(\eta_0\phi)|^2 dxdz + \gamma \int_{\Omega^\varepsilon} \eta_0^2 \{|\phi(x,0)|^2 + |\phi(x,l)|^2\} dx - \int_{D_l} |\nabla\eta_0|^2|\phi|^2 dxdz.$$

Since $|\nabla\eta_0| \leq C\varepsilon^{-1/3}$, we use (5.4) to get

$$S_1 \geq \left[\varepsilon^{-2} + \tau_\gamma(l)^2\right] \int_{D_l} |\eta_0\phi|^2 dxdz - O(\varepsilon^{-2/3}) \int_{D_l} |\phi|^2 dxdz.$$

Step 2. We estimate $\eta_1\phi$. The sum in (5.7) involving $\eta_1\phi$ is

$$S_2 \equiv \int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{F}}(\eta_1\phi)|^2 dxdz + \gamma \int_{\partial D_l} |\eta_1\phi|^2 dS - \int_{D_l} |\nabla\eta_1|^2|\phi|^2 dxdz.$$

Choose a family of cut-off functions $\{\chi_i\}_{i\in I}$ such that

$$\sum_{i\in I}\chi_i^2(x)=1 \quad \text{for all } x\in\Omega_{\varepsilon^{1/3}}, \qquad \sum_{i\in I}|\nabla\chi_i(x)|^2\le C\varepsilon^{-2/3}.$$

Then

$$\begin{aligned}S_2=&\sum_{i\in I}\left\{\int_{D_l}|\nabla_{\varepsilon^{-2}\mathbf{F}}(\chi_i\eta_1\phi)|^2dxdz+\gamma\int_{\partial D_l}|\chi_i\eta_1\phi|^2dS\right\}\\&-\int_{D_l}|\nabla\eta_1|^2|\phi|^2dxdz-\sum_{i\in I}\int_{D_l}|\nabla\chi_i|^2|\eta_1\phi|^2dxdz.\end{aligned}$$

For each $i\in I$, let $\phi_i=\chi_i\eta_1\phi$. Note that $\chi_i\eta_1$ depends only on x. $\bar D_l\cap\operatorname{spt}(\phi_i)$ is contained in a set G_i, whose coordinates (s,t,z) satisfy

$$|s-s_i|\le C\varepsilon^{1/3},\quad 0\le t\le\frac{\varepsilon^{1/3}}{10},\quad 0\le z\le l.$$

$\phi_i=0$ on three of the faces of G_i, which correspond with $t=\frac{\varepsilon^{1/3}}{10}$, $s=s_i-C\varepsilon^{1/3}$, and $s=s_i+C\varepsilon^{1/3}$, respectively. Hence, to estimate S_2, we are led to an eigenvalue problem on G_i with Dirichlet conditions on the three faces of G_i and Robin conditions on the other three faces of G_i. As in Theorem 5.1, we find the lowest eigenvalue λ_i of this problem by the method of separable variables:

$$\lambda_i=\frac{1}{\varepsilon^2}[\beta_0+C_1(3\gamma-\kappa_i)\varepsilon+O(\varepsilon^{4/3})]+\tau_\gamma(l)^2, \tag{5.8}$$

where

$$\kappa_i=\max_{x\in\partial\Omega\cap\bar G_i}\kappa_r(x).$$

Note that if $|s-s_i|\le C\varepsilon^{1/3}$, then $|\kappa_r(s)-\kappa_r(s_i)|\le C\varepsilon^{1/3}$. Using this and (5.8) we get

$$\begin{aligned}&\int_{D_l}|\nabla_{\varepsilon^{-2}\mathbf{F}}\phi_i|^2dxdz+\gamma\int_{\partial D_l}|\phi_i|^2dS\ge\lambda_i\int_{D_l}|\phi_i|^2dxdz\\&=\left\{\varepsilon^{-2}\left[\beta_0+C_1(3\gamma-\kappa_i)\varepsilon+O(\varepsilon^{4/3})\right]+\tau_\gamma(l)^2\right\}\int_{D_l}|\phi_i|^2dxdz\\&\ge\varepsilon^{-2}\int_{D_l}\left\{\beta_0+C_1(3\gamma-\kappa_r(s))\varepsilon+O(\varepsilon^{4/3})+\varepsilon^2\tau_\gamma(l)^2\right\}|\phi_i|^2dxdz.\end{aligned}$$

So

$$\begin{aligned}&\sum_{i\in I}\left\{\int_{D_l}|\nabla_{\varepsilon^{-2}\mathbf{F}}(\chi_i\eta_1\phi)|^2dxdz+\gamma\int_{\partial D_l}|\chi_i\eta_1\phi|^2dS\right\}\\&\ge\varepsilon^{-2}\int_{D_l}\left\{\beta_0+C_1(3\gamma-\kappa_r(s))\varepsilon+O(\varepsilon^{4/3})+\varepsilon^2\tau_\gamma(l)^2\right\}|\eta_1\phi|^2dxdz.\end{aligned}$$

From the choice of η_1 and χ_i we have

$$\begin{aligned}\int_{D_l}|\nabla\eta_1|^2|\phi|^2dxdz&\le C\varepsilon^{-2/3}\int_{D_l}|\phi|^2dxdz,\\\sum_{i\in I}\int_{D_l}|\nabla\chi_i|^2|\eta_1\phi|^2dxdz&\le C\varepsilon^{-2/3}\int_{D_l}|\phi|^2dxdz.\end{aligned}$$

Hence we have

$$S_2 \geq \varepsilon^{-2} \int_{D_l} \{\beta_0 + C_1(3\gamma - \kappa_r(s))\varepsilon + \varepsilon^2 \tau_\gamma(l)^2\} |\eta_1 \phi|^2 dxdz - O(\varepsilon^{-2/3}) \int_{D_l} |\phi|^2 dxdz.$$

Step 3. Combining the computations in Steps 1 and 2 we find

$$\begin{aligned}
&\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{F}} \phi|^2 dxdz + \gamma \int_{\partial D_l} |\phi|^2 dS = S_1 + S_2 \\
&\geq \varepsilon^{-2} \int_{D_l} \{[1 + \varepsilon^2 \tau_\gamma(l)^2] |\eta_0 \phi|^2 + [\beta_0 + C_1(3\gamma - \kappa_r(s))\varepsilon + \varepsilon^2 \tau_\gamma(l)^2] |\eta_1 \phi|^2\} dxdz \\
&\quad - O(\varepsilon^{-2/3}) \int_{D_l} |\phi|^2 dxdz \\
&\geq \varepsilon^{-2} \int_{D_l} \mathcal{W}_{D_l}(x,z) |\phi|^2 dxdz,
\end{aligned}$$

where $\mathcal{W}_{D_l}(x,z)$ is the function given in (5.6). □

COROLLARY 5.4. *Let* curl $\mathbf{F} = (0,0,1)$ *and let* ϕ_ε *be an eigenfunction associated with* $\mu_\gamma(\varepsilon^{-2}\mathbf{F})$. *Then there exist positive numbers* α_j, b_j, c_j, *and* M_j *independent of* l *such that, for all small* $\varepsilon > 0$,

(i) $\int_{D_l} \exp\big(\alpha_1 \varepsilon^{-1} d_{\partial\Omega}(x)\big) |\phi_\varepsilon|^2 dxdz \leq M_1 \int_{D_l \cap \{d_{\partial\Omega}(x) < c_1\varepsilon\}} |\phi_\varepsilon|^2 dxdz$;

(ii) $\int_{D_l} \exp\big(\alpha_2 \varepsilon^{-1/2} [\kappa_{\max} - \kappa_r(s) - b_2 \varepsilon^{1/3}]\big) |\phi_\varepsilon|^2 dxdz \leq M_2 \int_{D_l \cap \{d_{\partial\Omega}(x) < c_2\varepsilon\}} |\phi_\varepsilon|^2 dxdz$.

Moreover, if $l \gg \varepsilon$,

(iii) $\int_{D_l} \exp\big(\alpha_3 \varepsilon^{-1} d_l(x,z)\big) |\phi_\varepsilon|^2 dxdz \leq M_3 \int_{D_l \cap \{d_l(x,z) < c_3\varepsilon\}} |\phi_\varepsilon|^2 dxdz$.

Remark 4. The above corollary shows that, as $\varepsilon \to 0$, the eigenfunctions of (5.1) associated with the lowest eigenvalue localize in the strip $\mathcal{N}(\partial\Omega) \times [0,l]$. Conclusion (i) gives an exponential decay of the eigenfunctions in the direction normal to the lateral surface, conclusion (ii) shows an exponential decay in tangential direction away from $\mathcal{N}(\partial\Omega) \times [0,l]$, and conclusion (iii) shows an exponential decay away from the total boundary of D_l. These conclusions are proved using the Agmon argument [A], and the details are omitted (see the proof of Theorem 6.2 in section 6 for a related estimate).

6. Films in perpendicular fields. In this section we study a film with small l and large κ (l may vary with κ) that is placed in a perpendicular applied field, namely, $\mathbf{h} \equiv \mathbf{e}_3 = (0,0,1)$. We shall estimate the value of H_{C_3} and study nucleation of superconductivity. Let curl $\mathbf{F} = \mathbf{e}_3$.

First, using Theorem 5.1 and applying the argument in [LP4, Appendix], we obtain a lower bound of H_{C_3}.

LEMMA 6.1. *When* $l - 2\gamma\kappa^{-2} \gg \kappa^{-4}$ *we have*

$$H_{C_3}(\mathbf{e}_3, \kappa, l) \geq \left(1 - \frac{2\gamma}{l\kappa^2}\right) \frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}} (\kappa_{\max} - 3\gamma) \left(1 - \frac{2\gamma}{l\kappa^2}\right)^{1/2} + O(\kappa^{-1/3}).$$

Proof. As observed in [LP4],

$$H_{C_3} \geq \sigma_* \equiv \sigma_*(\kappa) = \max\{\sigma > 0 : \mu_\gamma(\kappa\sigma\mathbf{F}) = \kappa^2\}.$$

Using the equality

$$\kappa^2 = \mu_\gamma(\kappa\sigma_*\mathbf{F}) = \beta_\gamma(\kappa\sigma_*\mathbf{E}) + \tau_\gamma(l)^2,$$

where $\mathbf{E} = (-x_2, 0)$, and the condition $l - 2\gamma\kappa^{-2} \gg \kappa^{-4}$, we find that as $\kappa \to \infty$, $\beta_\gamma(\kappa\sigma_*\mathbf{E}) \to \infty$, and hence $\kappa\sigma_* \to \infty$. Then we use Theorem 5.1 to get

$$\kappa^2 = \beta_0\kappa\sigma_* + C_1(3\gamma - \kappa_{\max})\sqrt{\kappa\sigma_*} + O((\kappa\sigma_*)^{1/3}) + \frac{2\gamma}{l} - \frac{\gamma^2}{3} + O(\gamma^3 l).$$

Thus

$$\sqrt{\kappa\sigma_*} = \frac{1}{2\beta_0}\left\{C_1(\kappa_{\max} - 3\gamma) + \left[4\beta_0\left(\kappa^2 - \frac{2\gamma}{l}\right) + O(\kappa\sigma_*)^{1/3}\right]^{1/2}\right\}.$$

So $\sigma_* = O(\kappa)$, and

$$\sigma_* = \left(1 - \frac{2\gamma}{l\kappa^2}\right)\frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma)\left(1 - \frac{2\gamma}{l\kappa^2}\right)^{1/2} + O(\kappa^{-1/3}).$$

This yields the conclusion. □

The upper bound of H_{C_3} and the nature of nucleation of superconductivity vary according to the scale of l and κ. In Theorem 6.2 we consider the case where l has order of κ^{-1}, namely, there exist positive constants $a < b$ such that

$$a\kappa^{-1} \le l \le b\kappa^{-1}. \tag{6.1}$$

In Theorems 6.3, 6.4, and 6.5, we consider the case where l has order of κ^{-2}.

THEOREM 6.2. *Assume the condition* (6.1).

(i) *Given* $0 < \alpha < 1/3$ *we have, for large* κ,

$$H_{C_3}(\mathbf{e}_3, \kappa, l) = \frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma) - \frac{2\gamma}{\beta_0\kappa l} + o(\kappa^{-\alpha}). \tag{6.2}$$

(ii) *As the applied field decreases from* $H_{C_3}(\mathbf{e}_3, \kappa, l)$, *superconductivity nucleates first in the strip* $\mathcal{N}(\partial\Omega) \times [0, l]$. *More precisely, assume that*

$$\sigma = H_{C_3}(\mathbf{e}_3, \kappa, l) - \rho > 0. \tag{6.3}$$

For any $0 < p < 1/9$, *there exist positive constants* c_1, c_2, *and* C *such that for any minimal solution* $(\psi, \mathbf{A})$ *of* (1.1), *we have*

$$\int_{D_l} \exp\bigl(c_1\sqrt{\kappa}[\kappa_{\max} - \kappa_r(x) - c_2\rho - c_2\kappa^{-p}]\bigr)|\psi|^2 dxdz \le C\kappa^{-2}. \tag{6.4}$$

Proof. Step 1. Lemma 6.1 provides a lower bound of $H_{C_3}(\mathbf{e}_3, \kappa, l)$. To prove (6.2), let us fix $0 < \alpha < 1/3$ and choose σ such that

$$\frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma) - \frac{2\gamma}{\beta_0\kappa l} + O(\kappa^{-1/3}) < \sigma < H_{C_3}(\mathbf{e}_3, \kappa, l). \tag{6.5}$$

For our convenience we introduce $\varepsilon = (\kappa\sigma)^{-1/2}$. From (6.1) and (6.5) we see that l has order of ε. From Lemma 6.1, the Ginzburg–Landau functional has nontrivial minimizers, which will be denoted by $(\psi^\varepsilon, \mathbf{A}^\varepsilon)$. We need some elliptic estimates for $(\psi^\varepsilon, \mathbf{A}^\varepsilon)$. In the following, C denotes a generic constant which may vary from line to line.

CLAIM 1. *Assume* (6.1) *and* (6.5).

(i) *There exists* $C_1(a,b)>0$ *independent of* l *and* ε *such that*

$$\begin{aligned} &\|\psi^\varepsilon\|_{L^\infty(D_l)} = o(1) \quad \text{as } \varepsilon\to 0,\\ &\|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}\psi^\varepsilon\|_{L^\infty(D_l)} \le \frac{C_1(a,b)}{\varepsilon}\|\psi^\varepsilon\|_{L^\infty(D_l)}. \end{aligned} \tag{6.6}$$

(ii) *For any* $0<\alpha<1$, *there exists* $C_2(a,b,\alpha)>0$ *independent of* l *and* ε, *and* $q=3/(1-\alpha)$ *such that*

$$\|\mathbf{A}^\varepsilon-\mathbf{F}\|_{C^{1+\alpha}(\bar D_l)} \le \varepsilon C_2(a,b,\alpha)\big\{\|\psi^\varepsilon\|^2_{L^4(D_l)} + \|\psi^\varepsilon\|_{L^q(D_l)}\|\psi^\varepsilon\|_{L^\infty(D_l)}\big\}. \tag{6.7}$$

(iii) *For any* $0<\alpha<2\sqrt{1-\beta_0}$, *there exists* $C_3(a,b,\alpha)>0$ *independent of* l *and* ε *such that*

$$\int_{D_l}\exp\big(\alpha\varepsilon^{-1}d_{\partial\Omega}(x)\big)\{|\psi^\varepsilon|^2+\varepsilon^2|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}\psi^\varepsilon|^2\}dxdz \le C_3(a,b,\alpha)l\varepsilon. \tag{6.8}$$

As in the proof of Lemma 4.1 (Step 4), we use a blow-up argument in the scale ε to prove (6.6). Since l has order of ε, the blow-up process leads to a limiting equation with bounded coefficients in a domain of the form $(-\infty,\infty)\times(0,\infty)\times(0,L)$, where $L>0$. We first apply the L^∞ estimates to this equation and then return to the original scale and obtain the second inequality in (6.6). To prove the first inequality in (6.6), we use the fact that the limiting nonlinear equation has no nontrivial bounded solutions (see [LP4, Proposition 2.5 and Theorem 5.1] for two-dimensional problems).

Equation (6.7) follows from (6.6) and (4.8).

Equation (6.8) can be proved by an Agmon-type argument [A]. For a two-dimensional linear problem, such an estimate has been established by Helffer and Morame [HM1, (6.25), (6.59)]. For a two-dimensional nonlinear problem, a similar estimate was proved in [P2, Lemma 7.2]. Here we sketch an outline of the proof. Let χ be a smooth function which vanishes at $\partial\Omega\times[0,l]$. From the equation for ψ^ε we get

$$\int_{D_l}|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}(\chi\psi^\varepsilon)|^2dxdz+\gamma\int_{\partial D_l}|\chi\psi^\varepsilon|^2dS = \int_{D_l}\big\{\kappa^2(1-|\psi^\varepsilon|^2)|\chi\psi^\varepsilon|^2+|\nabla\chi|^2|\psi^\varepsilon|^2\big\}dxdz. \tag{6.9}$$

From (6.7) we have $|\mathrm{curl}(\mathbf{A}^\varepsilon-\mathbf{F})|=O(\varepsilon)$. Thus

$$\partial_1 A_2^\varepsilon-\partial_2 A_1^\varepsilon \ge 1+O(\varepsilon).$$

So we use Lemma 5.2 to get

$$\int_{D_l}|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}(\chi\psi^\varepsilon)|^2dxdz+\gamma\int_{\partial D_l}|\chi\psi^\varepsilon|^2dS \ge \frac{1+O(\varepsilon)+\varepsilon^2\tau_\gamma(l)^2}{\varepsilon^2}\int_{D_l}|\chi\psi^\varepsilon|^2dxdz.$$

From this and (6.9), and using the fact $\varepsilon^2\kappa^2=\frac{\kappa}{\sigma}=\beta_0+o(1)<1$ (see (6.5)), we have

$$\int_{D_l}|\chi\psi^\varepsilon|^2dxdz \le \frac{\varepsilon^2(1+o(1))}{1-\beta_0}\int_{D_l}|\nabla\chi|^2|\psi^\varepsilon|^2dxdz.$$

Now we choose

$$\chi(x,z)=\chi(x)=\eta(x)\exp\Big(\frac{\alpha}{2}\varepsilon^{-1}\zeta(x)\Big),$$

where $0 < \alpha < 2\sqrt{1-\beta_0}$ is a constant, $\zeta(x)$ is a smooth function on $\bar{\Omega}$ such that $\zeta(x) = d_{\partial\Omega}(x)$ on Ω_{δ_0}, $\eta(x)$ is a cut-off function such that $\eta(x) = 0$ if $d_{\partial\Omega}(x) < \varepsilon$, $\eta(x) = 1$ if $d_{\partial\Omega}(x) > 2\varepsilon$, and $|\nabla\eta(x)| \le \frac{4}{\varepsilon}$. Plugging it into the above inequality we find

$$\int_{D_l} \eta^2 \exp\bigl(\alpha\varepsilon^{-1}\zeta\bigr)|\psi^\varepsilon|^2 dxdz \le C\int_{D_l\cap\{\mathrm{dist}(x,\partial\Omega)<2\varepsilon\}} |\psi^\varepsilon|^2 dxdz \le Cl\varepsilon,$$

which implies that for a larger C,

$$\int_{D_l} \exp\bigl(\alpha\varepsilon^{-1}d_{\partial\Omega}(x)\bigr)|\psi^\varepsilon|^2 dxdz \le Cl\varepsilon.$$

Using this and (6.9) we find

$$\int_{D_l} \exp\bigl(\alpha\varepsilon^{-1}d_{\partial\Omega}(x)\bigr)|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}\psi^\varepsilon|^2 dxdz \le \frac{Cl}{\varepsilon}.$$

Hence (6.8) is true. Now Claim 1 is proved.

As a consequence of (6.8) we have, for any nonnegative integer k,

$$\int_{D_l} d_{\partial\Omega}(x)^k\bigl\{|\psi^\varepsilon|^2 + \varepsilon^2|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}\psi^\varepsilon|^2\bigr\}dxdz \le C_k l\varepsilon^{k+1}. \tag{6.10}$$

In particular

$$\int_{D_l} |\psi^\varepsilon|^2 dxdz \le Cl\varepsilon. \tag{6.11}$$

We define, for $q > 3$,

$$d(\varepsilon) = \|\psi^\varepsilon\|^2_{L^4(D_l)} + \|\psi^\varepsilon\|_{L^q(D_l)}\|\psi^\varepsilon\|_{L^\infty(D_l)}. \tag{6.12}$$

From (6.11) we find

$$d(\varepsilon) \le C\bigl\{(l\varepsilon)^{1/2}\|\psi^\varepsilon\|_{L^\infty(D_l)} + (l\varepsilon)^{1/q}\|\psi^\varepsilon\|^{2(q-1)/q}_{L^\infty(D_l)}\bigr\}. \tag{6.13}$$

Step 2. Now we establish the weighted L^2 estimates. In the following, let ϕ^ε denote the eigenfunction of the operator $-\nabla^2_{\varepsilon^{-2}\mathbf{A}^\varepsilon}$ associated with the lowest eigenvalue $\mu_\gamma(\varepsilon^{-2}\mathbf{A}^\varepsilon)$. Similar to Corollary 5.4 we have the following.

Claim 2. *Assume* (6.1) *and* (6.5).

(i) *For any* $0 < \alpha < 2\sqrt{1-\beta_0}$, *there exists a positive constant* $C_4(a,b,\alpha)$ *independent of* l *and* ε *such that*

$$\int_{D_l} \exp\bigl(\alpha\varepsilon^{-1}d_{\partial\Omega}(x,z)\bigr)\{|\phi^\varepsilon|^2 + \varepsilon^2|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}\phi^\varepsilon|^2\}dxdz \le \varepsilon C_4(a,b,\alpha)\int_{D_l}|\phi^\varepsilon|^2 dxdz. \tag{6.14}$$

(ii) *There exists a positive constant* C *independent of* l *and* ε *such that for any smooth function* χ *we have*

$$\begin{aligned}&\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}(\chi\phi^\varepsilon)|^2 dxdz + \gamma\int_{\partial D_l}|\chi\phi^\varepsilon|^2 dS\\ &\quad\ge \varepsilon^{-2}\int_{D_l}\mathcal{W}_{D_l}(x,z)|\chi\phi^\varepsilon|^2 dxdz - C\varepsilon^{-4/3}d(\varepsilon)^{2/3}\int_{D_l}|\chi\phi^\varepsilon|^2 dxdz.\end{aligned} \tag{6.15}$$

Here $\mathcal{W}_{D_l}(x,z)$ is the function given in (5.6).

Equation (6.14) can be proved as for (6.8). Before proving (6.15), we would like to mention that in the two-dimensional case studied in [HP], the magnetic potentials vanish at the boundary. This condition makes it easy to control error terms in various estimates. In the three-dimensional case discussed here, on the contrary, this condition is no longer true, and hence a harder estimate is needed in order to obtain (6.15). We should also mention the difference between (5.5) and (6.15), although they look similar. The inequality (5.5) is regarding a fixed vector field, $\mathbf{F}$, while (6.15) gives a uniform estimate for a family of vector fields $\mathbf{A}^\varepsilon$. When proving (5.5), we divided D_l into subsets in the form of the product $\Omega_j \times [0,l]$, where $\{\Omega_j\}$ is a partition of Ω. In order to prove (6.15), we need a finer partition and have to choose the cut-off functions carefully. If l is small compared with ε (for example, if l has order of ε), we may choose cut-off functions to be independent of z, and hence cut D_l into a finite number of blocks with height l. If l is not very small, the interval $[0,l]$ should also be divided, and hence the cut-off functions must depend on z. In the following we describe the choice of cut-off functions which are valid for a general case, not only for the purpose here.

Let $\{\eta_{0,\tau(\varepsilon)}, \eta_{1,\tau(\varepsilon)}\}$ be the partition of unity on $\mathbb{R}$ introduced in [HM2, (9.22), (9.23)] such that

$$\eta^2_{0,\tau(\varepsilon)}(t) + \eta^2_{1,\tau(\varepsilon)}(t) = 1, \qquad |\eta'_{j,\tau(\varepsilon)}(t)| \le \frac{C}{\tau(\varepsilon)} \quad \text{for } j = 0,1,$$
$$\text{spt } (\eta_{0,\tau(\varepsilon)}) \subset \left[\frac{\tau(\varepsilon)}{20}, +\infty\right), \qquad \text{spt } (\eta_{1,\tau(\varepsilon)}) \subset \left(-\infty, \frac{\tau(\varepsilon)}{10}\right].$$

As in [HM1, (9.10)–(9.14)], let $\{\chi_i(x,z)\}_I$ be a partition of unity of $\mathbb{R}^3$ such that $I = \mathbb{Z}^3$ and

$$\chi_i \in C^\infty(\mathbb{R}^3, \mathbb{R}) \quad \text{and} \quad \text{spt } (\chi_i) \subset i + [-1,1]^3 \quad \text{for any } i \in I,$$
$$\sum_{i\in I} \chi_i^2(x,z) = 1, \qquad \sum_{i\in I} |\nabla \chi_i(x,z)|^2 < C.$$

Let $\delta_0 > 0$ be chosen such that the distance function $d_{\partial\Omega}$ is differentiable in Ω_{δ_0}. Let $\omega(\varepsilon)$ be a function of ε such that $0 < \omega(\varepsilon) < \delta_0$ for all small ε, and set

$$\chi_{i,\omega(\varepsilon)}(x,z) = \chi_i\left(\frac{x}{\omega(\varepsilon)}, \frac{z}{\omega(\varepsilon)}\right), \qquad i \in I.$$

Thus we have a new partition of unity such that

$$\begin{aligned} &\text{spt}(\chi_{i,\omega(\varepsilon)}) \subset \omega(\varepsilon) i + [-\omega(\varepsilon), \omega(\varepsilon)]^3, \\ &\sum_{i\in I} \chi_{i,\omega(\varepsilon)}(x,z)^2 = 1, \qquad \sum_{i\in I} |\nabla \chi_{i,\omega(\varepsilon)}(x,z)|^2 < \frac{C}{\omega(\varepsilon)^2}. \end{aligned} \tag{6.16}$$

In the following, we choose $0 < \tau(\varepsilon) \le \omega(\varepsilon)$. Let us introduce

$$I(\omega(\varepsilon)) = \big\{ i \in I \ : \ \text{spt}(\chi_{i,\omega(\varepsilon)}) \cap D_l \neq \emptyset, \quad \text{dist}\big(\text{spt}(\chi_{i,\omega(\varepsilon)}), \partial\Omega \times [0,l]\big) \le \omega(\varepsilon) \big\}.$$

Note that Ω_{δ_0} can be covered by a finite number of open sets Ω_j such that on each of them we can define a diffeomorphism that straightens a portion of $\partial\Omega$. For simplicity we let f denote any of these maps. Then $\Omega_{\delta_0} \times [0,l]$ is covered by a finite number of

open sets $\Omega_j \times [a_k, b_k]$, and on each of them we can define a diffeomorphism of the form $\mathcal{F}(y,z) = (f(y), z)$. For each $i \in I(\omega(\varepsilon))$, we can choose $x(i) \in \partial\Omega$, $z(i) \in [0,l]$ such that

$$\operatorname{spt}(\chi_{i,\varepsilon}) \subset B^2\left(x(i), \frac{3}{2}\omega(\varepsilon)\right) \times [z(i) - \omega(\varepsilon), z(i) + \omega(\varepsilon)],$$

where $B^2\big(x(i), \frac{3}{2}\omega(\varepsilon)\big)$ denotes a two-dimensional disc with center $x(i)$ and radius $\frac{3}{2}\omega(\varepsilon)$. Choose $y(i) = (y_1(i), 0)$ such that $\mathcal{F}(y(i), z(i)) = (x(i), z(i))$. Let

$$K(i,\varepsilon) = \{\mathcal{F}(y_1, y_2, z) :\ |y_1 - y_1(i)| < 2\omega(\varepsilon),\ 0 \le y_2 \le \tau(\varepsilon),\ |z - z(i)| \le 2\omega(\varepsilon)\}.$$

Then, for small $\varepsilon > 0$, $\operatorname{spt}(\chi_{i,\varepsilon}) \cap \bar{D}_l \subseteq K(i,\varepsilon)$, and $\{K(i,\varepsilon) : i \in I(\omega(\varepsilon))\}$ covers the thin cylinder $\Omega_{\tau(\varepsilon)} \times [0,l]$.

Given a smooth function χ, we write

$$\begin{aligned}
&\phi_\varepsilon(x,z) = \chi\phi^\varepsilon(x,z),\\
&u_\varepsilon(x,z) = \eta_{1,\tau(\varepsilon)}(t(x))\phi_\varepsilon(x,z),\\
&u_{i,\varepsilon}(x,z) = \chi_{i,\omega(\varepsilon)}u_\varepsilon(x,z),
\end{aligned}$$

where $t(x) = d_{\partial\Omega}(x)$. Note that for any smooth functions ϕ and ψ we have

(6.17)
$$\begin{aligned}
&\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}\phi|^2 dxdz + \gamma\int_{\partial D_l} |\phi|^2 dS\\
&= \sum_{j=0}^{1}\left\{\int_{D_l}\big[|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}(\eta_{j,\tau(\varepsilon)}\phi)|^2 - |\phi\nabla\eta_{j,\tau(\varepsilon)}|^2\big]dxdz + \gamma\int_{\partial D_l}|\eta_j\phi|^2 dS\right\},\\
&\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}\psi|^2 dxdz + \gamma\int_{\partial D_l} |\psi|^2 dS\\
&= \sum_{i\in I(\omega(\varepsilon))}\left\{\int_{D_l}\big[|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}(\chi_{i,\omega(\varepsilon)}\psi)|^2 - |\psi\nabla\chi_{i,\omega(\varepsilon)}|^2\big]dxdz + \gamma\int_{\partial D_l}|\psi\chi_{i,\omega(\varepsilon)}|^2 dS\right\}.
\end{aligned}$$

Applying the second equality in (6.17) to $\psi = u_\varepsilon$, and using (6.16), we find

(6.18)
$$\begin{aligned}
&\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}u_\varepsilon|^2 dxdz + \gamma\int_{\partial D_l} |u_\varepsilon|^2 dS\\
&= \sum_{i\in I(\omega(\varepsilon))}\left\{\int_{D_l}|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}u_{i,\varepsilon}|^2 dxdz + \gamma\int_{\partial D_l}|u_{i,\varepsilon}|^2 dS\right\} - \frac{O(1)}{\omega(\varepsilon)^2}\int_{D_l}|u_\varepsilon|^2 dxdz,
\end{aligned}$$

where $O(1)$ remains bounded as $\varepsilon \to 0$.

Now we fix $i \in I(\omega(\varepsilon))$ and estimate the integral of $|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}u_{i,\varepsilon}|^2$. Note that

$$\int_{D_l}|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}u_{i,\varepsilon}|^2 dxdz = \int_{K(i,\varepsilon)}|\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}u_{i,\varepsilon}|^2 dxdz.$$

On $K(i,\varepsilon)$ we write

$$\mathbf{A}^\varepsilon(x,z) = \mathbf{F}(x,z) + \mathbf{B}_i(x,z) + \mathbf{b}_i, \qquad u_{i,\varepsilon}(x,z) = v_{i,\varepsilon}(x,z)e^{i\mathbf{b}_i\cdot(x,z)},$$

where $\mathbf{b}_i = \mathbf{A}^\varepsilon(x(i), z(i)) - \mathbf{F}(x(i), z(i))$. From (6.7) we have

$$\|\mathbf{B}_i\|_{C^{1+\alpha}(\bar{D}_l)} = O(\varepsilon)d(\varepsilon), \qquad |\mathbf{b}_i| = O(\varepsilon),$$

where $d(\varepsilon)$ is given in (6.12) with $q = 3/(1-\alpha)$. Since $\mathbf{B}_i(y(i), z(i)) = \mathbf{0}$, we have

$$|\mathbf{B}_i(x,z)| \le C\varepsilon d(\varepsilon)\sqrt{|x-x(i)|^2 + |z-z(i)|^2} \le C\varepsilon d(\varepsilon)\omega(\varepsilon) \qquad \text{on } K(i,\varepsilon). \tag{6.19}$$

We compute

$$\begin{aligned}
&\int_{K(i,\varepsilon)} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon} u_{i,\varepsilon}|^2 dxdz = \int_{K(i,\varepsilon)} |\nabla_{\varepsilon^{-2}(\mathbf{F}+\mathbf{B}_i)} v_{i,\varepsilon}|^2 dxdz \\
&= \int_{K(i,\varepsilon)} \big\{ |\nabla_{\varepsilon^{-2}\mathbf{F}} v_{i,\varepsilon}|^2 - 2\varepsilon^{-2}\mathbf{B}_i \cdot \Im(\bar{v}_{i,\varepsilon}\nabla_{\varepsilon^{-2}\mathbf{F}} v_{i,\varepsilon}) + \varepsilon^{-4}|\mathbf{B}_i|^2 |v_{i,\varepsilon}|^2 \big\} dxdz \\
&\ge \int_{K(i,\varepsilon)} |\nabla_{\varepsilon^{-2}\mathbf{F}} v_{i,\varepsilon}|^2 dxdz - 2\varepsilon^{-2} \int_{K(i,\varepsilon)} \mathbf{B}_i \cdot \Im(\bar{v}_{i,\varepsilon}\nabla_{\varepsilon^{-2}\mathbf{F}} v_{i,\varepsilon}) dxdz.
\end{aligned}$$

Using (6.19) we have, for a constant $M > 0$,

$$\begin{aligned}
&2\varepsilon^{-2}\Big| \int_{K(i,\varepsilon)} \mathbf{B}_i \cdot \Im(\bar{v}_{i,\varepsilon}\nabla_{\varepsilon^{-2}\mathbf{F}} v_{i,\varepsilon}) dxdz \Big| \\
&\le M\varepsilon^{-4} \int_{K(i,\varepsilon)} |\mathbf{B}_i|^2 |v_{i,\varepsilon}|^2 dxdz + M^{-1} \int_{K(i,\varepsilon)} |\nabla_{\varepsilon^{-2}\mathbf{F}} v_{i,\varepsilon}|^2 dxdz \\
&\le CMd(\varepsilon)^2\omega(\varepsilon)^2\varepsilon^{-2} \int_{K(i,\varepsilon)} |v_{i,\varepsilon}|^2 dxdz + M^{-1} \int_{K(i,\varepsilon)} |\nabla_{\varepsilon^{-2}\mathbf{F}} v_{i,\varepsilon}|^2 dxdz.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\int_{K(i,\varepsilon)} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon} u_{i,\varepsilon}|^2 dxdz \\
&\ge (1 - M^{-1}) \int_{K(i,\varepsilon)} |\nabla_{\varepsilon^{-2}\mathbf{F}} v_{i,\varepsilon}|^2 dxdz - CMd(\varepsilon)^2\omega(\varepsilon)^2\varepsilon^{-2} \int_{K(i,\varepsilon)} |v_{i,\varepsilon}|^2 dxdz.
\end{aligned}$$

Since $|v_{i,\varepsilon}| = |u_{i,\varepsilon}|$, using Theorem 5.3 and the above inequality we find

$$\begin{aligned}
&\int_{K(i,\varepsilon)} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon} u_{i,\varepsilon}|^2 dxdz + \gamma \int_{\partial D_l} |u_{i,\varepsilon}|^2 dS \\
&\ge (1 - M^{-1})\varepsilon^{-2} \int_{K(i,\varepsilon)} \mathcal{W}_{D_l} |u_{i,\varepsilon}|^2 dxdz - CMd(\varepsilon)^2\omega(\varepsilon)^2\varepsilon^{-2} \int_{K(i,\varepsilon)} |u_{i,\varepsilon}|^2 dxdz.
\end{aligned} \tag{6.20}$$

From (6.18) and (6.20) we get

$$\begin{aligned}
&\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon} u_\varepsilon|^2 dxdz + \gamma \int_{\partial D_l} |u_\varepsilon|^2 dS \\
&\ge (1 - M^{-1})\varepsilon^{-2} \sum_{i \in I(\omega(\varepsilon))} \int_{K(i,\varepsilon)} \mathcal{W}_{D_l} |u_{i,\varepsilon}|^2 dxdz \\
&\quad - CMd(\varepsilon)^2\omega(\varepsilon)^2\varepsilon^{-2} \sum_{i \in I(\omega(\varepsilon))} \int_{K(i,\varepsilon)} |u_{i,\varepsilon}|^2 dxdz - \frac{O(1)}{\omega(\varepsilon)^2} \int_{D_l} |u_\varepsilon|^2 dxdz \\
&\ge (1 - M^{-1})\varepsilon^{-2} \int_{D_l} \mathcal{W}_{D_l} |u_\varepsilon|^2 dxdz - C\left\{ Md(\varepsilon)^2\omega(\varepsilon)^2\varepsilon^{-2} + \frac{1}{\omega(\varepsilon)^2} \right\} \int_{D_l} |u_\varepsilon|^2 dxdz.
\end{aligned}$$

Equation (6.1) implies that $\mathcal{W}_{D_l}$ is uniformly bounded. If we choose

$$\omega(\varepsilon) = \varepsilon^{1/2} d(\varepsilon)^{-1/2} M^{-1/4},$$

then we get

$$\begin{aligned}
&\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon} u_\varepsilon|^2 dxdz + \gamma \int_{\partial D_l} |u_\varepsilon|^2 dS \\
&\geq (1 - M^{-1})\varepsilon^{-2} \int_{D_l} \mathcal{W}_{D_l} |u_\varepsilon|^2 dxdz - Cd(\varepsilon) M^{1/2} \varepsilon^{-1} \int_{D_l} |u_\varepsilon|^2 dxdz \\
&\geq \varepsilon^{-2} \int_{D_l} \mathcal{W}_{D_l} |u_\varepsilon|^2 dxdz - C\{M^{-1}\varepsilon^{-2} + d(\varepsilon) M^{1/2} \varepsilon^{-1}\} \int_{D_l} |u_\varepsilon|^2 dxdz.
\end{aligned}$$

Now we let

$$M = [\varepsilon d(\varepsilon)]^{-2/3}.$$

Then

$$\omega(\varepsilon) = \varepsilon^{2/3} d(\varepsilon)^{-1/3},$$

and we obtain

$$\begin{aligned}
&\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon} u_\varepsilon|^2 dxdz + \gamma \int_{\partial D_l} |u_\varepsilon|^2 dS \\
&\geq \varepsilon^{-2} \int_{D_l} \mathcal{W}_{D_l} |u_\varepsilon|^2 dxdz - C\varepsilon^{-4/3} d(\varepsilon)^{2/3} \int_{D_l} |u_\varepsilon|^2 dxdz.
\end{aligned} \tag{6.21}$$

Next, we choose $\tau(\varepsilon) = \omega(\varepsilon) = \varepsilon^{2/3} d(\varepsilon)^{-1/3}$ and use (6.21) and the first equality in (6.17) to get

(6.22)

$$\begin{aligned}
&\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon} (\chi\phi^\varepsilon)|^2 dxdz + \gamma \int_{\partial D_l} |\chi\phi^\varepsilon|^2 dS \\
&\geq \int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon} u_\varepsilon|^2 dxdz + \gamma \int_{\partial D_l} |u_\varepsilon|^2 dS - C\varepsilon^{-4/3} d(\varepsilon)^{2/3} \int_{D_l} |\chi\phi^\varepsilon|^2 dxdz \\
&\geq \varepsilon^{-2} \int_{D_l} \mathcal{W}_{D_l} |u_\varepsilon|^2 dxdz - C\varepsilon^{-4/3} d(\varepsilon)^{2/3} \int_{D_l} |u_\varepsilon|^2 dxdz - C\varepsilon^{-4/3} d(\varepsilon)^{2/3} \int_{D_l} |\chi\phi^\varepsilon|^2 dxdz.
\end{aligned}$$

Finally, using (6.14), we find

$$\begin{aligned}
\int_{D_l} \mathcal{W}_{D_l} |u_\varepsilon|^2 dxdz &= \int_{D_l} [\mathcal{W}_{D_l} + O(\varepsilon)] |\chi\phi^\varepsilon|^2 dxdz, \\
\int_{D_l} |u_\varepsilon|^2 dxdz &= \int_{D_l} [1 + O(\varepsilon)] |\chi\phi^\varepsilon|^2 dxdz.
\end{aligned}$$

Using these equalities and (6.22) we get (6.15).

Step 3. We establish a uniform lower bound estimate of the lowest eigenvalue. From (6.15) we have, for all small $\varepsilon > 0$,

$$\mu_\gamma(\varepsilon^{-2}\mathbf{A}^\varepsilon) \geq \frac{1}{\varepsilon^2}\{\beta_0 + C_1(3\gamma - \kappa_{\max})\varepsilon + \varepsilon^2 \tau_\gamma(l)^2 - C\varepsilon^{4/3} - C\varepsilon^{2/3} d(\varepsilon)^{2/3}\}.$$

From (6.6) and (6.13) we have $d(\varepsilon) = o((l\varepsilon)^{1/q})$. Thus

$$\mu_\gamma(\varepsilon^{-2}\mathbf{A}^\varepsilon) \geq \frac{1}{\varepsilon^2}\Big\{\beta_0 + C_1(3\gamma - \kappa_{\max})\varepsilon + \varepsilon^2\tau_\gamma(l)^2 - C\varepsilon^{4/3} - o(l^{\frac{2}{3q}}\varepsilon^{\frac{2(1+q)}{3q}})\Big\}. \tag{6.23}$$

Note that (6.23) holds if we only assume $l \geq c\kappa^{-1}$.

Now we use the condition (6.1) to simplify (6.23) and get

$$\mu_\gamma(\varepsilon^{-2}\mathbf{A}^\varepsilon) \geq \frac{1}{\varepsilon^2}\Big\{\beta_0 + C_1(3\gamma - \kappa_{\max})\varepsilon + \varepsilon^2\tau_\gamma(l)^2 - C\varepsilon^{4/3} - o(\varepsilon^b)\Big\},$$

where $b = \frac{2(2+q)}{3q}$. For any $0 < \alpha < 1/3$ we can choose $3 < q < 4$ such that $b = 1 + \alpha$. So, under the condition (6.1) we have

$$\mu_\gamma(\varepsilon^{-2}\mathbf{A}^\varepsilon) \geq \frac{1}{\varepsilon^2}\Big\{\beta_0 + C_1(3\gamma - \kappa_{\max})\varepsilon + \varepsilon^2\tau_\gamma(l)^2 - o(\varepsilon^{1+\alpha})\Big\}. \tag{6.24}$$

Step 4. We prove an upper bound estimate of $H_{C_3}(\mathbf{e}_3, \kappa, l)$. Recall that in order to have nontrivial minimizers we must have $\mu_\gamma(\varepsilon^{-2}\mathbf{A}^\varepsilon) < \kappa^2$; see [LP4]. Using this and (6.1), (6.24) we find that

$$\sigma \leq \frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma) - \frac{\tau_\gamma(l)^2}{\beta_0\kappa} + o(\kappa^{-\alpha}).$$

From this and (2.5) we get

$$H_{C_3}(\mathbf{e}_3, \kappa, l) \leq \frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma) - \frac{2\gamma}{\beta_0 l\kappa} + o(\kappa^{-\alpha}).$$

Now (6.2) is proved.

Step 5. Using (6.2), the conclusion about the location of nucleation can be proved following the argument in [HP, proof of Theorem 6.1]. In the following we outline the proof. Assume σ satisfies (6.3). Again we let $\varepsilon = (\kappa\sigma)^{-1/2}$, and let $(\psi^\varepsilon, \mathbf{A}^\varepsilon)$ be the minimizer. For any smooth function χ, we have (6.9). As in Step 2 we have (see (6.22))

$$\begin{aligned}&\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}(\chi\psi^\varepsilon)|^2 dxdz + \gamma\int_{\partial D_l} |\chi\psi^\varepsilon|^2 dS\\ &\geq \varepsilon^{-2}\int_{D_l} \mathcal{W}_{D_l}|\chi\psi^\varepsilon|^2 dxdz - C\varepsilon^{-4/3}d(\varepsilon)^{2/3}\int_{D_l} |\chi\psi^\varepsilon|^2 dxdz.\end{aligned}$$

Using this and (6.9) we get

$$\int_{D_l} |\chi\psi^\varepsilon|^2\{\mathcal{W}_{D_l} - \varepsilon^2\kappa^2 - C\varepsilon^{2/3}d(\varepsilon)^{2/3}\}dxdz \leq \varepsilon^2\int_{D_l} |\nabla\chi|^2|\psi^\varepsilon|^2 dxdz. \tag{6.25}$$

From (6.2) and (6.3) we have, for some α with $p < \alpha < 1/3$,

$$\begin{aligned}\varepsilon^2\kappa^2 &= \frac{\kappa}{\sigma} = \beta_0 - \left[\sqrt{\beta_0}C_1(\kappa_{\max} - 3\gamma) - \frac{2\gamma\beta_0}{\kappa l}\right]\kappa^{-1} + \beta_0^2\rho\kappa^{-1} + o(\kappa^{-1-\alpha}),\\ \varepsilon &= \frac{1}{\sqrt{\kappa\sigma}} = \sqrt{\beta_0}\kappa^{-1} + O(\kappa^{-2}).\end{aligned}$$

So we have

$$\varepsilon^2\mu_\gamma(\varepsilon^{-2}\mathbf{F})-\varepsilon^2\kappa^2=\beta_0+C_1(3\gamma-\kappa_{\max})\varepsilon+\varepsilon^2\tau_\gamma(l)^2-\varepsilon^2\kappa^2+O(\varepsilon^{4/3})$$
$$=\frac{2\gamma}{l\kappa}\left(\frac{1}{\sigma}-\sqrt{\beta_0}\varepsilon\right)-\beta_0^{3/2}\varepsilon\rho+O(\varepsilon^{1+\alpha})=-\beta_0^{3/2}\varepsilon\rho+O(\varepsilon^{1+\alpha}).$$

From this and (6.25) we get

$$\varepsilon^2\int_{D_l}|\nabla\chi|^2|\psi^\varepsilon|^2dxdz$$
$$\geq\int_{D_l}|\chi\psi^\varepsilon|^2\{\mathcal{W}_{D_l}-\varepsilon^2\mu_\gamma(\varepsilon^{-2}\mathbf{F})-\beta_0^{3/2}\varepsilon\rho-C\varepsilon^{2/3}d(\varepsilon)^{2/3}+O(\varepsilon^{1+\alpha})\}dxdz.$$

Using Theorems 5.1 and 5.3 we find

$$\int_{D_l}\big\{C_1[k_{\max}-\kappa_r(x)]-\beta_0^{3/2}\rho-C\varepsilon^\alpha-C\varepsilon^{-1/3}d(\varepsilon)^{2/3}\big\}|\chi\psi^\varepsilon|^2dxdz\leq\varepsilon\int_{D_l}|\nabla\chi|^2|\psi^\varepsilon|^2dxdz.$$

Recall that $d(\varepsilon)=o(l\varepsilon)^{1/q}$ for some $q>3$. Thus $\varepsilon^{-1/3}d(\varepsilon)^{2/3}=o(\varepsilon^b)$, where $b=\frac{4}{3q}-\frac{1}{3}$. We can choose q close to 3 such that $b\geq\alpha>p$.

Let $\phi(x)$ be a smooth function such that $\phi(x)=\kappa_{\max}-\kappa_r(x)$ in Ω_{δ_0}. There exists $C_0>0$ such that $|\nabla\phi|^2\leq C_0\phi$. Let $a=\sqrt{\frac{C_1}{2C_0}}$ and

$$\chi=\exp\big(a\varepsilon^{-1/2}\phi(x)\big).$$

Plugging it into the above integral inequality we get

$$\int_{D_l}|\psi^\varepsilon|^2\exp\big(2a\varepsilon^{-1/2}\phi\big)\big\{C_1[k_{\max}-\kappa_r(x)]-2\beta_0^{3/2}\rho-o(\varepsilon^p)\big\}dxdz\leq 0.$$

So there exist positive constants c and M_1 such that

$$\int_{\{\kappa_{\max}-\kappa_r(x)\geq c\rho+c\varepsilon^p,\ 0\leq z\leq l\}}|\psi^\varepsilon|^2\exp\big(2a\varepsilon^{-1/2}(\phi-c\rho-c\varepsilon^p)\big)dxdz$$
$$\leq M_1\int_{\{\kappa_{\max}-\kappa_r(x)\leq c\rho+c\varepsilon^p,\ 0\leq z\leq l\}}|\psi^\varepsilon|^2dxdz,$$

and hence

$$\int_{D_l}|\psi^\varepsilon|^2\exp\big(2a\varepsilon^{-1/2}(\phi-c\rho-c\varepsilon^p)\big)dxdz\leq 2M_1\int_{\{\kappa_{\max}-\kappa_r(x)\leq c\rho+c\varepsilon^p,\ 0\leq z\leq l\}}|\psi^\varepsilon|^2dxdz.$$

Using this and (6.8) we get, for some $M_3>M_2>2M_1$ and $m>0$,

$$\int_{D_l}|\psi^\varepsilon|^2\exp\big(2a\varepsilon^{-1/2}(\phi-c\rho-c\varepsilon^p)\big)dxdz$$
$$\leq M_2\int_{\{\kappa_{\max}-\kappa_r(x)\leq c\varepsilon^p,\ d_{\partial\Omega}(x)<m\varepsilon,\ 0\leq z\leq l\}}|\psi^\varepsilon|^2dxdz\leq M_3l\varepsilon.$$

From this we get (6.4). □

Remark 5. (i) From (6.23) and Lemma 6.1 we find that if $l \geq a\kappa^{-1}$, then for any $q > 3$ fixed, we have

$$H_{C_3}(\mathbf{e}_3, \kappa, l) = \frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma) - \frac{2\gamma}{\beta_0 l\kappa} + o(l^{\frac{2}{3q}}\kappa^{\frac{q-2}{3q}}). \tag{6.26}$$

In particular, if there exists α_0, $1/2 < \alpha_0 < 1$, such that $a\kappa^{-1} \leq l \leq b\kappa^{-\alpha_0}$, then we choose $3 < q < 2 + 2\alpha_0$ in (6.26) and also get (6.2) with $\alpha = \alpha_0$.

(ii) We expect that if $l \gg \kappa^{-1}$, the solutions on D_l behave as the solutions on a bulk domain, say, a cylinder with a finite height,[11] and we expect that superconductivity nucleates at a subset of the top and bottom edges, namely, at the set

$$\mathcal{N}(\partial\Omega) \times \{0\} \bigcup \mathcal{N}(\partial\Omega) \times \{l\}.$$

In the following we consider the films with $l \sim a\kappa^{-2}$. We shall see in Theorems 6.3, 6.4, and 6.5 that $H_{C_3}(\mathbf{e}_3, \kappa, l)$ depends sensitively on the value of a, and $a = 2\gamma$ is a critical value.

THEOREM 6.3. *Assume that*

$$l = a\kappa^{-2} + O(\kappa^{-4}), \qquad a > 2\gamma. \tag{6.27}$$

(i) *Let* $\frac{1}{6} < \alpha < \frac{1}{3}$. *For large* κ *we have*

$$H_{C_3}(\mathbf{e}_3, \kappa, l) = \left(1 - \frac{2\gamma}{a}\right)\frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma)\left(1 - \frac{2\gamma}{a}\right)^{1/2} + o(\kappa^{-\alpha}). \tag{6.28}$$

(ii) *As the applied field decreases from* $H_{C_3}(\mathbf{e}_3, \kappa, l)$, *superconductivity nucleates first in the strip* $\mathcal{N}(\partial\Omega) \times [0, l]$. *More precisely, assume that* σ *satisfies* (6.3), *and let* $(\psi, \mathbf{A})$ *be the minimizer of the Ginzburg–Landau functional* $\mathcal{G}$. *Then for any* $0 < p < 1/3$, *there exist positive constants* c_1, c_2, *and* C *such that* (6.4) *holds.*

Proof. We modify the proof of Theorem 6.2 to get the conclusions.

Step 1. From Lemma 6.1 we get a lower bound of H_{C_3}:

$$H_{C_3}(\mathbf{e}_3, \kappa, l) \geq \left(1 - \frac{2\gamma}{a}\right)\frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma)\left(1 - \frac{2\gamma}{a}\right)^{1/2} + O(\kappa^{-1/3}). \tag{6.29}$$

Now let us choose σ such that

$$\left(1 - \frac{2\gamma}{a}\right)\frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma)\left(1 - \frac{2\gamma}{a}\right)^{1/2} + O(\kappa^{-1/3}) < \sigma < H_{C_3}(\mathbf{e}_3, \kappa, l). \tag{6.30}$$

Again we let $\varepsilon = (\kappa\sigma)^{-1/2}$. Then l has order of ε^2. Let $(\psi^\varepsilon, \mathbf{A}^\varepsilon)$ be a minimizer of the Ginzburg–Landau functional. Then, as in the proof of Theorem 6.2, we have $\|\psi^\varepsilon\|_{L^\infty(D_l)} = o(1)$. Using the second inequality of (4.9) we have, for any $q > 3$,

$$\|\Im\{\bar{\psi}^\varepsilon \nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}\psi^\varepsilon\}\|_{L^q(D_l)} \leq \frac{C(q)}{\varepsilon}\|\psi^\varepsilon\|_{L^q(D_l)}\|\psi^\varepsilon\|_{L^\infty(D_l)}. \tag{6.31}$$

[11] Minimal solutions on such domains have been examined in [P1].

We use (4.8) to get

$$\|\mathbf{A}^\varepsilon - \mathbf{F}\|_{C^{1+\alpha}(\bar{D}_l)} \le C(\alpha)\varepsilon\{\|\psi^\varepsilon\|^2_{L^4(D_l)} + \|\psi^\varepsilon\|_{L^q(D_l)}\|\psi^\varepsilon\|_{L^\infty(D_l)}\}.$$

Then we use the argument in the proof of Theorem 6.2 to show that (see (6.11))

$$\|\psi^\varepsilon\|_{L^2(D_l)} = O((l\varepsilon)^{1/2}).$$

Hence for $0 < \alpha_0 < 1$ and $q = 3/(1-\alpha_0)$,

$$\|\mathbf{A}^\varepsilon - \mathbf{F}\|_{C^{1+\alpha_0}(\bar{D}_l)} = o(\varepsilon)d(\varepsilon), \qquad d(\varepsilon) = (l\varepsilon)^{1/2} + (l\varepsilon)^{1/q} = O(\varepsilon^{\frac{3}{q}}),$$

since $l = O(\varepsilon^2)$. Let ϕ^ε be the eigenfunction of $-\nabla^2_{\varepsilon^{-2}\mathbf{A}^\varepsilon}$ associated with the lowest eigenvalue. As in the proof of Theorem 6.2 (Step 2) we can show that for any smooth function χ,

(6.32)
$$\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}(\chi\phi^\varepsilon)|^2 dxdz \ge \int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{F}}(\chi\phi^\varepsilon)|^2 dxdz - C\varepsilon^{-4/3}d(\varepsilon)^{2/3}\int_{D_l}|\chi\phi^\varepsilon|^2 dxdz.$$

Then, similar to (6.23), now we have

$$\mu_\gamma(\varepsilon^{-2}\mathbf{A}^\varepsilon) \ge \frac{1}{\varepsilon^2}\{\beta_0 + C_1(3\gamma - \kappa_{\max})\varepsilon + \varepsilon^2\tau_\gamma(l)^2 - o(\varepsilon^{\frac{2}{3}+\frac{2}{q}})\}.$$

Since $(\psi^\varepsilon, \mathbf{A}^\varepsilon)$ is nontrivial, we have $\mu_\gamma(\varepsilon^{-2}\mathbf{A}^\varepsilon) < \kappa^2$. If $\frac{1}{6} < \alpha < \frac{1}{3}$, we choose $3 < q < 4$ such that $\frac{2}{q} - \frac{1}{3} = \alpha$. Using this and (6.32) we find

$$H_{C_3}(\mathbf{e}_3, \kappa, l) \le \left(1 - \frac{2\gamma}{a}\right)\frac{\kappa}{\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma)\left(1 - \frac{2\gamma}{a}\right)^{1/2} + o(\kappa^{-\alpha}).$$

Equation (6.28) follows from this and (6.29).

Step 2. Using the argument in the proof of Theorem 6.2 (Step 5) and (6.28), we can show that superconductivity nucleates in the strip $\mathcal{N}(\partial\Omega)\times[0,l]$. In fact, now (6.25) also holds, and $d(\varepsilon) = O(\varepsilon^{3/q})$. From (6.28) and (6.3) we have, for some α with $p < \alpha < 1/3$,

$$\begin{aligned}
\frac{\kappa}{\sigma} &= \beta_0\left(1 - \frac{2\gamma}{a}\right)^{-1} - C_1(\kappa_{\max} - 3\gamma)\frac{\sqrt{\beta_0}}{\kappa}\left(1 - \frac{2\gamma}{a}\right)^{-3/2} \\
&\quad + \frac{\beta_0^2\rho}{\kappa}\left(1 - \frac{2\gamma}{a}\right)^{-2} + o(\kappa^{-1-\alpha}), \\
\varepsilon &= \frac{\sqrt{\beta_0}}{\kappa}\left(1 - \frac{2\gamma}{a}\right)^{-1/2} + O(\kappa^{-2}), \\
\varepsilon^2\mu_\gamma(\varepsilon^{-2}\mathbf{F}) - \varepsilon^2\kappa^2 &= -\beta_0^{3/2}\left(1 - \frac{2\gamma}{a}\right)^{-1/2}\varepsilon\rho + o(\varepsilon^{1+\alpha}).
\end{aligned}$$

Thus we have

$$\begin{aligned}
&\int_{D_l}\left\{[C_1(\kappa_{\max} - \kappa_r(x)] - \beta_0^{3/2}\left(1 - \frac{2\gamma}{a}\right)^{-1/2}\rho - o(\varepsilon^\alpha)\right\}|\chi\psi^\varepsilon|^2 dxdz \\
&\quad \le \varepsilon\int_{D_l}|\nabla\chi|^2|\psi^\varepsilon|^2 dxdz.
\end{aligned}$$

Then we can proceed as in Step 5 of the proof of Theorem 6.2 to obtain (6.4). □

Next we consider the case where

$$l = 2\gamma\kappa^{-2} + c\kappa^{-3} + O(\kappa^{-4}), \qquad c > 0. \tag{6.33}$$

We shall see that $H_{C_3}(\mathbf{e}_3, \kappa, l)$ remains bounded between two positive numbers. The discussion is valid for a more general case $2\gamma\kappa^{-2} + a\kappa^{-3} \le l \le 2\gamma\kappa^{-2} + b\kappa^{-3}$.

THEOREM 6.4. *Assume* (6.33).

(i) *For large* κ *we have*

$$H_{C_3}(\mathbf{e}_3, \kappa, l) = \frac{c}{2\gamma\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma)\sqrt{\frac{c}{2\gamma}}\kappa^{-1/2} + O(\kappa^{-2/3}). \tag{6.34}$$

(ii) *Fix* σ_0 *such that*

$$\frac{c}{2\gamma} < \sigma_0 < \frac{c}{2\gamma\beta_0}, \tag{6.35}$$

and let $(\psi, \mathbf{A})$ *be the minimizer of the Ginzburg–Landau functional* $\mathcal{G}$ *for* $\sigma = \sigma_0$. *Then for any constant* $0 < \alpha < 2\sqrt{\sigma_0 - \frac{c}{2\gamma}}$ *there exists* $C(\alpha) > 0$ *independent of* κ *and* l *such that*

$$\int_{D_l} \exp\bigl(\alpha\sqrt{\kappa}d_{\partial\Omega}(x)\bigr)\{|\psi|^2 + \kappa^{-2}|\nabla_{\sigma_0\kappa\mathbf{A}}\psi|^2\}dxdz \le C\kappa^{-5/2}. \tag{6.36}$$

Proof. Step 1. From (6.33) and (2.5) we have

$$\kappa^2 - \tau_\gamma(l)^2 = \frac{c}{2\gamma}\kappa + O(1). \tag{6.37}$$

Using (6.37) and the proof of Lemma 6.1 we get the lower bound of H_{C_3}. To obtain an upper bound of H_{C_3}, we choose σ such that

$$\sigma_* + O(\kappa^{-1/3}) < \sigma < H_{C_3}(\mathbf{e}_3, \kappa, l).$$

Note that σ is bounded away from 0. Let $\varepsilon = (\kappa\sigma)^{-1/2}$ and let $(\psi^\varepsilon, \mathbf{A}^\varepsilon)$ be the minimizer. From the second inequality in (4.9) we have, for $q > 3$,

$$\|\Im\{\bar{\psi}^\varepsilon\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}\psi^\varepsilon\}\|_{L^q(D_l)} \le \frac{C(q)}{\varepsilon}\|\psi^\varepsilon\|_{L^q(D_l)}\|\psi^\varepsilon\|_{L^\infty(D_l)}.$$

For $q = 3/(1-\alpha)$, where $0 < \alpha < 1$, we get

$$\|\mathbf{A}^\varepsilon - \mathbf{F}\|_{C^{1+\alpha}(\bar{D}_l)} \le C\delta, \qquad \text{where} \qquad \delta = \frac{1}{\sigma}\|\psi^\varepsilon\|_{L^4(D_l)}^2 + \varepsilon\|\psi^\varepsilon\|_{L^q(D_l)}\|\psi^\varepsilon\|_{L^\infty(D_l)}. \tag{6.38}$$

Note that $\|\psi\|_{L^q(D_l)} = O(l^{1/q})$. Hence

$$\delta \le C(l^{1/2} + \varepsilon l^{1/q}).$$

Let ϕ^ε be the eigenfunction of $-\nabla^2_{\varepsilon^{-2}\mathbf{A}^\varepsilon}$ associated with the lowest eigenvalue. As in the proof of Theorem 6.2 (Step 2), we find

$$\int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{A}^\varepsilon}\phi^\varepsilon|^2 dxdz \ge \int_{D_l} |\nabla_{\varepsilon^{-2}\mathbf{F}}\phi^\varepsilon|^2 dxdz - C\varepsilon^{-2}\delta^{2/3}\int_{D_l} |\phi^\varepsilon|^2 dxdz. \tag{6.39}$$

Therefore

$$\kappa^2 > \mu_\gamma(\varepsilon^{-2}\mathbf{A}^\varepsilon) \geq \tau_\gamma(l)^2 + \frac{1}{\varepsilon^2}\Big\{\beta_0 + C_1(3\gamma - \kappa_{\max})\varepsilon - C\varepsilon^{4/3} - C\delta^{2/3}\Big\}. \tag{6.40}$$

So

$$\kappa^2 - \tau_\gamma(l)^2 > \frac{\beta_0 + o(1)}{\varepsilon^2} = [\beta_0 + o(1)]\kappa\sigma.$$

Using this and (6.37) we find

$$\sigma \leq \frac{c}{2\gamma\beta_0} + o(1).$$

Hence ε has order of $\kappa^{-1/2}$, and l has order of ε^4. We choose $3 < q < 4$ in (6.38) and find

$$\delta \leq C\varepsilon^2.$$

Then we write (6.40) as

$$\kappa^2 > \mu_\gamma(\varepsilon^{-2}\mathbf{A}^\varepsilon) \geq \tau_\gamma(l)^2 + \frac{1}{\varepsilon^2}\{\beta_0 + C_1(3\gamma - \kappa_{\max})\varepsilon - C\varepsilon^{4/3}\}.$$

From this we obtain

$$H_{C_3}(\mathbf{e}_3, \kappa, l) \leq \frac{c}{2\gamma\beta_0} + \frac{C_1}{\beta_0^{3/2}}(\kappa_{\max} - 3\gamma)\sqrt{\frac{c}{2\gamma}}\kappa^{-1/2} + O(\kappa^{-2/3}).$$

Equation (6.34) is proved.

Step 2. Let σ satisfy (6.35), and let $(\psi, \mathbf{A})$ be the minimizer for $\sigma = \sigma_0$. Let χ be a smooth function vanishing at $\partial\Omega \times [0, l]$. From the equation of ψ we get

$$\int_{D_l} |\nabla_{\kappa\sigma_0\mathbf{A}}(\chi\psi)|^2 dxdz + \gamma\int_{\partial D_l} |\chi\psi|^2 dS = \int_{D_l} \{\kappa^2(1 - |\psi|^2)|\chi\psi|^2 + |\nabla\chi|^2|\psi|^2\} dxdz. \tag{6.41}$$

Note that $|\psi| \leq 1$, and (6.38) remains true. We choose $3 < q < 4$ in (6.38) to find

$$\|\mathbf{A} - \mathbf{F}\|_{C^{1+\alpha}(\bar{D}_l)} \leq C\big\{\|\psi\|^2_{L^4(D_l)} + \kappa^{-1/2}\|\psi\|_{L^q(D_l)}\big\} \leq C(l^{1/2} + \kappa^{-1/2}l^{1/q}) \leq C\kappa^{-1}.$$

Thus $\partial_1\mathbf{A}_2 - \partial_2\mathbf{A}_1 = 1 + O(\kappa^{-1})$. Using this and Lemma 5.2 we get

$$\int_{D_l} |\nabla_{\kappa\sigma_0\mathbf{A}}(\chi\psi)|^2 dxdz + \gamma\int_{\partial D_l} |\chi\psi|^2 dS \geq \big[\kappa\sigma_0(1 + O(\kappa^{-1})) + \tau_\gamma(l)^2\big]\int_{D_l} |\chi\psi|^2 dxdz.$$

From this, (6.37), and (6.41) we have

$$\left\{\left(\sigma_0 - \frac{c}{2\gamma}\right)\kappa + O(1)\right\}\int_{D_l} |\chi\psi|^2 dxdz \leq \int_{D_l} |\nabla\chi|^2|\psi|^2 dxdz.$$

Now we choose

$$\chi(x, z) = \chi(x) = \eta(x)\exp\left(\frac{\alpha}{2}\sqrt{\kappa}\zeta\right),$$

where $0 < \alpha < 2\sqrt{\sigma_0 - \frac{c}{2\gamma}}$, $\zeta(x)$ is a smooth function on $\bar{\Omega}$ such that $\zeta(x) = d_{\partial\Omega}(x)$ on Ω_{δ_0}, $\eta(x)$ is a cut-off function such that $\eta(x) = 0$ if $d_{\partial\Omega}(x) < \kappa^{-1/2}$, $\eta(x) = 1$ if $d_{\partial\Omega}(x) > 2\kappa^{-1/2}$, and $|\nabla\eta(x)| \le 4\sqrt{\kappa}$. Plugging it into the above inequality we find

$$\int_{D_l} |\psi|^2 \eta^2 \exp(\alpha\sqrt{\kappa}\zeta) dxdz \le C \int_{D_l \cap \{\mathrm{dist}(x,\partial\Omega) < 2\kappa^{-1/2}\}} |\psi|^2 dxdz \le Cl\kappa^{-1/2},$$

which implies that for a larger C,

$$\int_{D_l} |\psi|^2 \exp\bigl(\alpha\sqrt{\kappa} d_{\partial\Omega}(x)\bigr) dxdz \le Cl\kappa^{-1/2}.$$

Using this and (6.41) we find

$$\int_{D_l} |\nabla_{\kappa\sigma_0\mathbf{A}}\psi|^2 \exp\bigl(\alpha\sqrt{\kappa} d_{\partial\Omega}(x)\bigr) dxdz \le Cl\kappa^{3/2}.$$

From these two inequalities and the condition (6.33) we get (6.36). □

THEOREM 6.5. *Assume that*

$$l(\kappa) < l < 2\gamma\kappa^{-2} + O(\kappa^{-4}), \tag{6.42}$$

where $l(\kappa)$ was defined in section 2. For large κ we have $H_{C_3}(\mathbf{e}_3, \kappa, l) = O(\kappa^{-1})$.

Proof. We show that $\kappa H_{C_3}(\mathbf{e}_3, \kappa, l)$ is bounded as $\kappa \to \infty$. Suppose this is not the case. Then for each κ we can find σ, $0 < \sigma < H_{C_3}(\mathbf{e}_3, \kappa, l)$, such that $\sigma\kappa \to \infty$ as $\kappa \to \infty$. Thus $\varepsilon = (\kappa\sigma)^{-1} \to 0$. Let $(\psi^\varepsilon, \mathbf{A}^\varepsilon)$ be the minimizer of the Ginzburg–Landau functional. As in the proof of Theorem 6.4 we have

$$\|\mathbf{A}^\varepsilon - \mathbf{F}\|_{C^{1+\alpha}(D_l)} \le C\delta, \qquad \text{where we choose} \qquad \delta = \frac{1}{\sigma}\|\psi^\varepsilon\|^2_{L^4(D_l)} + \varepsilon\|\psi^\varepsilon\|_{L^q(D_l)}.$$

Since $\sigma = \frac{1}{\varepsilon^2\kappa}$ and $\|\psi\|_{L^q(D_l)} = O(l^{1/q})$, we have

$$\delta \le C\varepsilon^2\kappa l^{1/2} + C\varepsilon l^{1/q} = O(\varepsilon^2 + \varepsilon l^{1/q}).$$

As in the proof of Theorem 6.4 we get (6.39). Hence, as $\varepsilon \to 0$,

$$\kappa^2 > \mu_\gamma(\varepsilon^{-2}\mathbf{A}^\varepsilon) - C\varepsilon^{-2}\delta^{2/3} = \tau_\gamma(l)^2 + \frac{\beta_0 + o(1)}{\varepsilon^2},$$

so

$$\beta_0 + o(1) \le \varepsilon^2[\kappa^2 - \tau_\gamma(l)^2] = \varepsilon^2\left[\frac{C}{2\gamma} + \frac{\gamma^2}{3} + o(1)\right] \to 0.$$

This contradiction shows that ε is bounded away from 0. Hence $\sigma \le C\kappa^{-1}$. □

Remark 6. Assume that $l = 2\gamma\kappa^{-2} + c\kappa^{-4}$. From Theorem 6.5, $H_{C_3}(\mathbf{e}_3, \kappa, l) \le C\kappa^{-1}$. Now we look for a lower bound. Let $a_0 > 0$ be the smallest positive number such that

$$\mu_\gamma(a_0\mathbf{E}) = \frac{c}{2\gamma} + \frac{\gamma^2}{3}.$$

Then

$$H_{C_3}(\mathbf{e}_3, \kappa, l) \ge \frac{a_0 + o(1)}{\kappa}. \tag{6.43}$$

To prove (6.43), we let $\sigma\kappa = a < a_0$ and choose test functions $(\psi, \mathbf{F})$ with $\psi(x,z) = a_l\xi_l(z)\phi(x)$, where ξ_l is the function given in (2.4), and $a_l = 1/\|\xi_l\|_{L^2([0,l])}$. Recall that $\mathbf{F} = (-x_2, 0, 0)$ and $\mathbf{E} = (-x_2, 0)$. We have

$$\mathcal{G}[\psi, \mathbf{F}] = l\mathcal{J}_\kappa[\phi] + \frac{\kappa^2 l|\Omega|}{2},$$

where

$$\mathcal{J}_\kappa[\phi] = \int_\Omega \left\{ |\nabla_{\kappa\sigma\mathbf{E}}\phi|^2 - (\kappa^2 - \tau_\gamma(l)^2)|\phi|^2 + \frac{\kappa^2 b_l}{2}|\phi|^4 \right\} dx + \gamma \int_{\partial\Omega} |\phi|^2 ds$$

and

$$b_l = \frac{\|\xi_l\|^4_{L^4([0,l])}}{\|\xi_l\|^4_{L^2([0,l])}}.$$

Note that $lb_l \to 1$ and $\kappa^2 b_l \to 2\gamma$ as $l \to 0$. When $\kappa\sigma = a < a_0$, the functional $\mathcal{J}_\kappa$ has a nontrivial minimizer with negative energy. Thus $\mathcal{G}$ has a nontrivial minimizer. So (6.43) is true.

Proof of Theorem 1.1. It follows from Theorems 6.2, 6.3, 6.4, and 6.5, and Remark 3. □

Remark 7. We may further discuss the behavior of thin films subjected to a perpendicular magnetic field far below H_{C_3}. We would like to give some observations.

(i) Consider a film with large κ and $l = 2\gamma\kappa^{-2} + c\kappa^{-3}$ placed in a perpendicular magnetic field $\mathcal{H} = \sigma\mathbf{e}_3$.

If $\frac{c}{2\gamma} < \sigma < \frac{c}{2\gamma\beta_0}$, superconductivity concentrates in a thin cylinder with thickness $O(\kappa^{-1/2})$ around the lateral surface $\partial\Omega \times [0, l]$. In fact, from Theorem 6.4, the order parameters exponentially decay in the direction normal to the lateral surface. Moreover the minimizers have energy

$$\mathcal{G}[\psi, \mathbf{A}] = \frac{\kappa^2 l|\Omega|}{2} + O(\kappa^{3/2} l).$$

If $0 < \sigma < \frac{c}{2\gamma}$, superconductivity occurs in the interior. To see this, let us compute the energy of the minimizers. Using (6.37) we can show that when $0 < \sigma < \frac{c}{2\gamma}$, the functional $\mathcal{J}_\kappa$ has a nontrivial minimizer $w_\kappa(x)$, and there exists a constant $C > 0$ such that for all large κ, $\int_\Omega |w_\kappa|^4 dx \geq C|\Omega|$ (see, for instance, [SS]). Hence

$$\mathcal{J}_\kappa[w_\kappa] = -\frac{\kappa^2 b}{2} \int_\Omega |w_\kappa|^4 dx \leq -\frac{C\kappa^2 l|\Omega|}{2}.$$

Let $\psi_\kappa = \alpha_l\xi_l(z)w_\kappa(x)$, and choose $(\psi_\kappa, \mathbf{F})$ as a test function. As in Remark 6 we find that $\mathcal{G}[\psi, \mathbf{F}] \leq \frac{1}{2}(1 - C)\kappa^2 l|\Omega|$. Thus the minimizers have energy

$$\mathcal{G}[\psi, \mathbf{A}] \leq \frac{1}{2}(1 - C)\kappa^2 l|\Omega|.$$

Hence the local energy of the minimizers in the interior is not negligible.

If we compare these phenomena with the behaviors of the minimizers of the two-dimensional Ginzburg–Landau functional (see [P2]), we may say that a film with thickness $l = 2\gamma\kappa^{-2} + c\kappa^{-3}$ exhibits a lateral surface superconductivity in a perpendicular field with magnitude $\frac{c}{2\gamma} < \sigma < \frac{c}{2\gamma\beta_0}$ and exhibits a bulk superconductivity if $0 < \sigma < \frac{c}{2\gamma\beta_0}$.

(ii) Now consider a film with $l = a\kappa^{-1}$, $a > 2\gamma$. From Theorems 6.2 and 6.3 we know that in a perpendicular magnetic field, superconductivity nucleates in a strip $\mathcal{N}(\partial\Omega) \times [0, l]$. Now consider the applied field far below H_{C_3} and assume $\mathcal{H} = \lambda\kappa\mathbf{e}_3$.

If $1 < \lambda < 1/\beta_0$, we expect that the minimizers concentrate at a thin layer around the lateral surface $\partial\Omega \times [0, l]$. Moreover, the behavior of the minimizers in the thin layer, after rescaling, can be described by the solutions of the limiting equation

$$\begin{cases} -\nabla^2_{\mathbf{F}}\psi = \lambda(1 - |\phi|^2)\psi & \text{in } \mathbb{R}^2_+ \times (0, a), \\ \nabla_{\mathbf{F}}\psi \cdot \nu = 0 & \text{if } x_2 = 0, \text{ or if } z = 0,\ a. \end{cases}$$

Thus we may say that in a perpendicular applied field lying in between κ and H_{C_3}, a film with thickness $l = a\kappa^{-1}$ behaves like a type II superconductor and exhibits a lateral surface superconducting state. We do not present the discussion here since it is similar to what we have done in [P2] for cylindrical domains with infinite height.[12]

If $0 < \lambda < 1$, we can show that superconductivity occurs at interior (see item (i)).

(iii) Similarly we can analyze the films with $l = a\kappa^{-2}$, where $a > 2\gamma$.

From Theorem 1.1 we have

$$H_{C_3}(\mathbf{e}_3, \kappa, l) \sim \begin{cases} \frac{\kappa}{\beta_0} & \text{if } l = a\kappa^{-1}, \quad a > 0, \\ (1 - \frac{2\gamma}{a})\frac{\kappa}{\beta_0} & \text{if } l = a\kappa^{-2}, \quad a > 2\gamma, \\ \frac{c}{2\gamma\beta_0} & \text{if } l = 2\gamma\kappa^{-2} + c\kappa^{-3}, \ \ c > 0. \end{cases}$$

The above observations suggest that

$$H_{C_2}(\mathbf{e}_3, \kappa, l) \sim \begin{cases} \kappa & \text{if } l = a\kappa^{-1}, \quad a > 0, \\ (1 - \frac{2\gamma}{a})\kappa & \text{if } l = a\kappa^{-2}, \quad a > 2\gamma, \\ \frac{c}{2\gamma} & \text{if } l = 2\gamma\kappa^{-2} + c\kappa^{-3}, \ \ c > 0. \end{cases}$$

REFERENCES

[A] S. Agmon, *Lectures on Exponential Decay of Solutions of Second Order Elliptic Equations: Bounds on Eigenfunctions of N-Body Schrödinger Operators*, Princeton University Press, Princeton, NJ, 1982.

[ADa] A. Aftalion and E. N. Dancer, *On the symmetry and uniqueness of solutions of the Ginzburg–Landau equations for small domains*, Commun. Contemp. Math., 3 (2001), pp. 1–14.

[ADu] A. Aftalion and Q. Du, *The bifurcation diagrams for the Ginzburg–Landau system of superconductivity*, Phys. D, 163 (2002), pp. 94–105.

[Al] Y. Almog, *Non-linear surface superconductivity for type* II *superconductors in the large domain limit*, Arch. Ration. Mech. Anal., 165 (2002), pp. 271–293.

[BPT] P. Bauman, D. Phillips, and Q. Tang, *Stable nucleation for the Ginzburg–Landau system with an applied magnetic field*, 142 (1998), pp. 1–43.

[BH] C. Bolley and B. Helffer, *An application of semi-classical analysis to the asymptotic study of the super cooling field of a superconducting material*, Ann. Inst. H. Poincaré Phys. Théor., 58 (1993), pp. 189–233.

[BR1] J. Berger and J. Rubinstein, *Formation of topological defects in thin superconducting rings*, Philos. Trans. Roy. Soc. London Ser. A, 355 (1997), pp. 1969–1978.

[BR2] J. Berger and J. Rubinstein, *Bifurcation analysis for phase transitions in superconducting rings with nonuniform thickness*, SIAM J. Appl. Math., 58 (1998), pp. 103–121.

[12] In [P2] we reduce the problem to two-dimensional domains, assuming that the solutions are invariant under translation in the direction of the axis of the cylinder. For a cylinder with finite height, this assumption should be dropped. However, the discussion in [P2] is still valid after a minor modification.

[BS] A. Bernoff and P. Sternberg, *Onset of superconductivity in decreasing fields for general domains*, J. Math. Phys., 39 (1998), pp. 1272–1284.

[C] S. J. Chapman, *Nucleation of superconductivity in decreasing fields*, I, II, European J. Appl. Math., 5 (1994), pp. 449–468, 469–494.

[CDG] S. Chapman, Q. Du, and M. Gunzburger, *A model for variable thickness superconducting thin films*, Z. Angew. Math. Phys., 47 (1996), pp. 410–431.

[CET] Z. Chen, C. Elliott, and Q. Tang, *Justification of a two dimensional evolutionary Ginzburg–Landau superconductivity model*, RAIRO Modél. Math. Anal. Numér., 32 (1998), pp. 25–50.

[CHO] S. J. Chapman, S. D. Howison, and J. R. Ockendon, *Macroscopic models for superconductivity*, SIAM Rev., 34 (1992), pp. 529–560.

[DD] S. J. Ding and Q. Du, *Critical magnetic field and asymptotic behavior of superconducting thin films*, SIAM J. Math. Anal., 34 (2002), pp. 239–256.

[DFS] M. del Pino, P. Felmer, and P. Sternberg, *Boundary concentration for eigenvalue problems related to the onset of superconductivity*, Comm. Math. Phys., 210 (2000), pp. 413–446.

[dG] P. G. de Gennes, *Superconductivity of Metals and Alloys*, W. A. Benjamin, New York, 1966.

[DG] Q. Du and M. Gunzburger, *A model for thin films having variable thickness*, Phys. D, 69 (1994), pp. 215–231.

[DGP] Q. Du, M. D. Gunzburger, and J. S. Peterson, *Analysis and approximation of the Ginzburg–Landau model of superconductivity*, SIAM Rev., 34 (1992), pp. 54–81.

[DH] M. Dauge and B. Helffer, *Eigenvalues variation,* I: *Neumann problem for Sturm-Liouville operators*, J. Differential Equations, 104 (1993), pp. 243–262.

[G] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations*, Vol. 1, Springer-Verlag, New York, 1994.

[GL] V. Ginzburg and L. Landau, *On the theory of superconductivity*, in Collected Papers, Gordon and Breach, New York, 1967, pp. 546–568.

[GP] T. Giorgi and D. Phillips, *The breakdown of superconductivity due to strong fields for the Ginzburg–Landau model*, SIAM J. Math. Anal., 30 (1999), pp. 341–359.

[GT] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1983.

[HM1] B. Helffer and A. Morame, *Magnetic bottles in connection with superconductivity*, J. Funct. Anal., 185 (2001), pp. 604–680.

[HM2] B. Helffer and A. Morame, *Magnetic bottles for the Neumann problem: The case of dimension* 3 *(general case)*, Proc. Indian Acad. Sci. Math. Sci., 112 (2002), pp. 71–84.

[HP] B. Helffer and X. B. Pan, *Upper critical field and location of surface nucleation of superconductivity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 145–181.

[HT] K.-H. Hoffmann and Q. Tang, *Ginzburg–Landau Phase Transition Theory and Superconductivity*, Internat. Ser. Numer. Math. 134, Birkhäuser-Verlag, Basel, 2001.

[J] H. Jadallah, *The onset of superconductivity in a domain with a corner*, J. Math. Phys., 42 (2001), pp. 4101–4121.

[JM] S. Jimbo and Y. Morita, *Ginzburg–Landau equation with magnetic effect in a thin domain*, Calc. Var. Partial Differential Equations, 15 (2002), pp. 325–352.

[L] O. A. Ladyzhenskaya, *The Boundary Value Problems of Mathematical Physics*, Springer-Verlag, New York, 1985.

[LP1] K. Lu and X. B. Pan, *Ginzburg–Landau equation with de Gennes boundary condition*, J. Differential Equations, 129 (1996), pp. 136–165.

[LP2] K. Lu and X. B. Pan, *Gauge invariant eigenvalue problems in* $\mathbb{R}^2$ *and in* $\mathbb{R}^2_+$, Trans. Amer. Math. Soc., 352 (2000), pp. 1247–1276.

[LP3] K. Lu and X. B. Pan, *Eigenvalue problems of Ginzburg–Landau operator in bounded domains*, J. Math. Phys., 40 (1999), pp. 2647–2670.

[LP4] K. Lu and X. B. Pan, *Estimates of the upper critical field for the Ginzburg–Landau equations of superconductivity*, Phys. D, 127 (1999), pp. 73–104.

[LP5] K. Lu and X. B. Pan, *Surface nucleation of superconductivity in* 3*-dimensions*, J. Differential Equations, 168 (2000), pp. 386–452.

[LP6] K. Lu and X. B. Pan, *Surface nucleation of superconductivity*, Methods Appl. Anal., 8 (2001), pp. 279–300.

[M] R. Montgomery, *Hearing the zero locus of a magnetic field*, Comm. Math. Phys., 168 (1995), pp. 651–675.

[P1] X. B. Pan, *Upper critical field for superconductors with edges and corners*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 447–482.

[P2] X. B. Pan, *Surface superconductivity in applied magnetic fields above* H_{C_2}, Comm. Math. Phys., 228 (2002), pp. 327–370.
[P3] X. B. Pan, *Superconductivity in* 3 *Dimensions*, preprint.
[P4] X. B. Pan, *Superconductivity near critical temperature*, J. Math. Phys., to appear.
[PK] X. B. Pan and K. H. Kwek, *Schrödinger operators with non-degenerately vanishing magnetic fields in bounded domains*, Trans. Amer. Math. Soc., 354 (2002), pp. 4201–4227.
[R] J. Rubinstein, *Six lectures on superconductivity*, in Boundaries, Interfaces, and Transitions, M. Delfour, ed., CRM Proc. Lecture Notes 13, AMS, Providence, RI, 1998, pp. 163–184.
[RR1] G. Richardson and J. Rubinstein, *A one-dimensional model for superconductivity in a thin wire of slowly varying cross-section*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 455 (1999), pp. 2549–2564.
[RR2] G. Richardson and J. Rubinstein, *The mixed boundary condition for the Ginzburg–Landau model in tin films*, Appl. Math. Lett., 13 (2000), pp. 97–99.
[RS] J. Rubinstein and M. Schatzman, *Asymptotics for thin superconducting rings*, J. Math. Pures Appl., 77 (1998), pp. 801–820.
[SdG] D. Saint-James and P. G. de Gennes, *Onset of superconductivity in decreasing fields*, Phys. Lett., 6 (1963), pp. 306–308.
[SS] E. Sandier and S. Serfaty, *The Decrease of Bulk-Superconductivity Close to the Second Critical Field in the Ginzburg–Landau Model*, preprint.
[SST] D. Saint-James, G. Sarma, and E. J. Thomas, *Type* II *Superconductivity*, Pergamon Press, Oxford, 1969.
[T] M. Tinkham, *Introduction to Superconductivity*, McGraw–Hill, New York, 1975.

SIAM J. MATH. ANAL.
Vol. 34, No. 4, pp. 992–1006

DROPLET SPREADING UNDER WEAK SLIPPAGE: A BASIC RESULT ON FINITE SPEED OF PROPAGATION*

GÜNTHER GRÜN†

Abstract. We prove a new qualitative result on finite speed of propagation for the thin film equation subjected to Navier slippage or even weaker slip conditions. Our approach works in multiple space dimensions and is based on a novel technique which combines recently established weighted energy estimates with a Hardy-type inequality and with Stampacchia's iteration lemma. It can be adapted to degenerate parabolic equations of order different from four as well.

Key words. fourth order degenerate parabolic equations, finite speed of propagation, thin films

AMS subject classifications. 35K35, 35K55, 35K65, 76D08

PII. S0036141002403298

1. Introduction. In this paper, we present a new method for the proof of qualitative results on finite speed of propagation for degenerate parabolic equations. It is inspired by the idea developed by Dal Passo, Giacomelli, and Grün in [10] to show the occurrence of a waiting time phenomenon for the thin film equation. We apply the method to establish new results on finite speed of propagation for the thin film equation

$$\begin{aligned} u_t + \operatorname{div}(|u|^n \nabla \Delta u) &= 0 \qquad \text{in } \mathbb{R}^N \times (0,\infty), \\ u(\cdot,0) &= u_0 \qquad \text{on } \mathbb{R}^N \end{aligned} \tag{1.1}$$

in the parameter range $n \in [2,3)$ in space dimensions $N < 4$. For ease of presentation, we confine ourselves to a mobility $m(u) := |u|^n$. However, modifications of (1.1) obtained by replacing the term $|u|^n$ with more general functions $m \in C^2(\mathbb{R};\mathbb{R}_0^+)$, $m(0) = 0$, could be handled as well.

Let us make a few comments on the physical background. In the course of lubrication approximation (see, e.g., Bernis [2] or Oron, Davis, and Bankoff [29]), the equation

$$h_t + \frac{\sigma}{3\eta} \operatorname{div}\left(m(h) \nabla \Delta h\right) = 0 \tag{1.2}$$

is derived to describe the surface tension driven evolution of the thickness h of a thin film of viscous liquid spreading on a horizontal surface. Here, η is the viscosity of the liquid and σ denotes surface tension. The explicit form of the mobility $m(\cdot)$ is determined by the flow condition at the liquid-solid interface. In case of a no-slip condition we get $m(h) = h^3$, whereas a generic slip condition

$$\vec{v}_{\text{hor}}\Big|_{z=0} = \beta h^{n-2} \cdot \frac{\partial \vec{v}_{\text{hor}}}{\partial z}\Big|_{z=0}$$

entails

$$m(h) = h^3 + \beta h^n.$$

*Received by the editors February 27, 2002; accepted for publication (in revised form) September 27, 2002; published electronically April 9, 2003.
http://www.siam.org/journals/sima/34-4/40329.html
†Institut für Angewandte Mathematik, Universität Bonn, Beringstr. 6, 53115 Bonn, Germany (gg@iam.uni-bonn.de).

Here, $\vec{v}_{\rm hor}$ denotes the horizontal component of the fluid velocity field, β is a positive parameter, and z stands for the vertical coordinate. For $n = 2$, the classical slip condition of Navier is recovered, but in the physical literature (cf., e.g., [15], [16], or [27]) different parameters of $n \in (0, 3)$ were suggested as well to model effects of stronger ($n < 2$) or weaker ($n \in (2, 3)$) slippage.

Analytically, the qualitative behavior of solutions is governed by the smoothness of $m(\cdot)$ in its point of degeneracy, i.e., in $h_0 = 0$. For this reason, in the mathematical literature usually the model problem (1.1) is studied, as it exhibits already all the essential mathematical difficulties. So let us emphasize that the physically important case of surface tension driven thin film flow subjected to Navier's slip condition or to even weaker slip conditions corresponds in the framework of the model problem (1.1) to the choices $n = 2$ or $2 < n < 3$, respectively. Hence, it is covered by the results to be presented in the present paper.

Recall that (1.1) admits globally nonnegative solutions (cf., e.g., Bernis and Friedman [5] or Grün [19]) and that it implicitly defines a free boundary problem where the free boundary at time $T \geq 0$ is given by $\partial[\mathrm{supp}(u(\cdot, T))]$. It is one of the striking features of (1.1) that the qualitative behavior of solutions is sensitive to the mobility growth exponent $n > 0$. Despite the fact that scaling invariances of (1.1) (for details cf. Giacomelli and Otto [14]) suggest the existence of compactly supported self-similar solutions for arbitrary $n > 0$, these solutions only exist for $0 < n < 3$ (cf. Bernis, Peletier, and Williams [6] and Ferreira and Bernis [13]). Moreover, for $n > 4$ the support of arbitrary solutions is constant in time, as was proven by Beretta, Bertsch, and Dal Passo [1]. And asymptotic analysis suggests that this behavior holds true for all $n \geq 3$. This is in good agreement with the spreading paradox observed by Dussan and Davis [12] in the framework of Navier–Stokes equations. It says that a no-slip condition at the liquid-solid interface (which entails $n = 3$) implies infinite energy dissipation at the liquid-solid-gas contact line in the case of moving droplets. On the other hand, it is well known (see Bernis and Friedman [5], Beretta, Bertsch, and Dal Passo [1], and Bertozzi and Pugh [7] in space dimension $N = 1$; Dal Passo, Garcke, and Grün [9] and Grün [18] in the multidimensional case) that for $n \in (0, 3)$ in space dimensions $N < 4$ so called *strong solutions* exist. They exhibit a zero contact angle at the free boundary, and for $t \to \infty$ their support tends to cover the spatial domain entirely.

The existence of compactly supported self-similar solutions indicates that solutions to (1.1) have the property of finite speed of propagation. In Bernis [3], [4] and in Bertsch et al. [8], this could be confirmed rigorously for $n \in (0, 2)$ in space dimensions $N < 4$ and for $n \in [2, 3)$ in space dimension $N = 1$. Surprisingly, the case $n \in [2, 3)$, $N > 1$, remained open for rather a long time. Only recently, the author succeeded in proving in his habilitation thesis [21] a first result on finite speed of propagation in the higher-dimensional setting for that critical parameter regime.

While the result of [21] took advantage of Bernis's technique of higher order differential inequalities, the new approach to be presented here essentially uses an iteration lemma due to Stampacchia, and it seems to be technically less involved. Let us emphasize that this method also applies to degenerate parabolic equations of sixth or even higher order or of second order like the *porous-media equation* or like *doubly degenerate parabolic equations* (cf., e.g., Vazquez [30] or Ivanov [26]).

Before describing the outline of the present paper, let us recall the peculiarities of the regime $n \in [2, 3)$ which seem to exclude an applicability of the techniques used for $n \in (0, 2)$. As (1.1) is fourth order parabolic, comparison principles do not hold and

the argumentation has to be based solely on integral estimates. At this time, there are basically two types of integral estimates known: first the energy estimate

$$\frac{1}{2}\int_{\mathbb{R}^N}|\nabla u(\cdot,T)|^2+\int_0^T\int_{\mathbb{R}^N}u^n|\nabla\Delta u|^2\le\frac{1}{2}\int_{\mathbb{R}^N}|\nabla u_0|^2 \tag{1.3}$$

and second the α-entropy estimate

$$\begin{aligned}\frac{1}{\alpha(\alpha+1)}\int_{\mathbb{R}^N}u^{\alpha+1}(\cdot,T)+C^{-1}\int_0^T\int_{\mathbb{R}^N}\Big\{|\nabla u^{\frac{\alpha+n+1}{4}}|^4+|D^2u^{\frac{\alpha+n+1}{2}}|^2\Big\}\\ \le\frac{1}{\alpha(\alpha+1)}\int_{\mathbb{R}^N}u_0^{\alpha+1},\end{aligned} \tag{1.4}$$

which is valid for

$$\alpha\in\left(\frac{1}{2}-n,2-n\right)\setminus\{-1,0\}. \tag{1.5}$$

Note that for compactly supported initial data, the global version of estimate (1.4) is valid only for $n\in(0,3)$. Observe, moreover, that condition (1.5) does not permit the parameter α to be chosen positive in the parameter regime $n\in[2,3)$. As a consequence, in that regime the entropy $u^{\alpha+1}(T)$ can no longer be controlled in terms of the initial entropy. This is the reason why analytical approaches based on the entropy estimate are restricted to the interval $n\in(0,2)$.

On the other hand, the energy estimate requires additional analytical tools in order to become accessible to Gagliardo–Nirenberg-type arguments. To this end, it would be desirable to estimate the dissipated energy $\int u^n|\nabla\Delta u|^2$ from below by derivatives of certain powers of u. In the multidimensional case, this goal was achieved only recently by virtue of the interpolation inequality

$$\int_\Omega|\nabla u^{\frac{n+2}{6}}|^6+\int_\Omega|\nabla\Delta u^{\frac{n+2}{2}}|^2\le C(n,N)\int_\Omega u^n|\nabla\Delta u|^2, \tag{1.6}$$

which was proven in [21] (see also the recently published paper [20]) and which holds on convex domains Ω for positive functions of class H^2 having zero normal derivatives on the boundary.

This result was the key observation to establish in [21] on bounded convex domains Ω the existence of strong solutions to the thin film equation associated with compactly supported, nonnegative initial data which satisfy, besides an α-entropy estimate, the following energy-type estimate:

$$\int_\Omega|\nabla u(\cdot,T)|^2+C(n,N)\left\{\int_0^T\int_\Omega|\nabla u^{\frac{n+2}{6}}|^6+\int_0^T\int_\Omega|\nabla\Delta u^{\frac{n+2}{2}}|^2\right\}\le\int_\Omega|\nabla u_0|^2. \tag{1.7}$$

By virtue of appropriate weighted versions of that estimate, first existence results for the Cauchy problem in the parameter regime $n\in(2-\sqrt{8/(8+N)},3)$ could be established as well; see [21].

In the present paper, we will prove that these solutions have the property of finite speed of propagation in the following sense.

DEFINITION 1.1. *Let $v:\mathbb{R}^N\times[0,\infty)\to\mathbb{R}$ be a nonnegative function and assume that $v(.,0)$ has compact support in $\mathbb{R}^N$. We say that v* has finite speed of propagation

iff for each ball $\overline{B_{R_0}(x_0)}$, $x_0 \in \mathbb{R}^N$, $R_0 > 0$, *that contains* $\operatorname{supp} v(\cdot,0)$, *a continuous, monotonically increasing function* $R : [0,\infty) \to \mathbb{R}_0^+$, $R(0) = 0$, *exists such that*

$$\operatorname{supp} v(.,t) \subset \overline{B_{R_0+R(t)}(x_0)}\,.$$

In section 2, we will recall the properties of strong solutions to the Cauchy problem as constructed in [21]. Section 3 is devoted to the proof of a Hardy-type inequality on exterior domains. Combining that Hardy-type inequality with the aforementioned weighted version of the energy estimate, we may formulate an integral estimate which will serve as the key ingredient for the subsequent proof of finite speed of propagation which follows in section 4. The idea of proof is to derive via appropriate interpolation arguments a recursive inequality for the function

$$G_T(R) := \int_0^T \left(\int_{\mathbb{R}^N \setminus B_R} u^2 \right)^{\frac{n+2}{2}}$$

which permits an application of Stampacchia's iteration lemma (see Lemma 4.3). This way, we deduce for fixed $T > 0$ the existence of a number $0 < R(T) < \infty$ such that $G_T(R(T)) = 0$. As a consequence, it becomes evident that $\operatorname{supp}(u(\cdot,t)) \subset B_{R(T)}(0)$ for all $0 \le t < T$. Furthermore, $R(T)$ continuously depends on T, which gives the result.

Throughout the paper, we use the usual notation for Sobolev and Lebesgue spaces. We write $\|u\|_p$ for $(\int |u|^p)^{1/p}$ also in the case $0 < p < 1$. Finally, $B_R(x)$ denotes the ball with radius R and center $x \in \mathbb{R}^N$, and $[u > 0]_T$ stands for the set $\{(x,t) \in \mathbb{R}^N \times (0,T) | u(x,t) > 0\}$.

2. Preliminaries. In this section, we will summarize recent results on strong solutions for the Cauchy problem in the multidimensional case. In addition, we will formulate a version of Gagliardo–Nirenberg's inequality to be used in what follows.

THEOREM 2.1. *Let* $n \in (2 - \sqrt{8/(8+N)}, 3)$, $N < 4$, *and assume* $u_0 \in H^1(\mathbb{R}^N)$ *to be nonnegative with compact support in the sense that* $u_0(x) = 0$ *almost everywhere on* $\mathbb{R}^N \setminus B_{R_0}(0)$ *for a positive number* R_0. *Then a nonnegative function* u *exists that has the following properties:*

(i) *Regularity:*

$$(2.1)\quad u_t \in L^2(\mathbb{R}^+; (W^{1,p}(\Omega))') \text{ for } p > \frac{4N}{2N + n(2-N)} \text{ and any } \Omega \subset\subset \mathbb{R}^N,$$

$$(2.2)\quad u \in L^\infty(\mathbb{R}^+; H^1(\mathbb{R}^N)),$$

$$(2.3)\quad D^2 u^{\frac{\alpha+n+1}{2}} \in L^2(\mathbb{R}^N \times \mathbb{R}^+) \text{ for any } \alpha \in (\max\{-1, 1/2-n\}, 2-n),$$

$$(2.4)\quad \nabla u^{\frac{\alpha+n+1}{4}} \in L^4(\mathbb{R}^N \times \mathbb{R}^+) \text{ for any } \alpha \in (\max\{-1, 1/2-n\}, 2-n),$$

$$(2.5)\quad J = \begin{cases} u^n \nabla \Delta u & \text{on } [u>0]_T \\ 0 & \text{on } [u=0]_T \end{cases} \in L^2(\mathbb{R}^+; L^q(\mathbb{R}^N))$$

$$\text{for any } 1 < q < \frac{4N}{2N + n(N-2)}.$$

(ii) u *is a solution to the Cauchy problem in the sense that*

$$(2.6)\qquad \int_0^T \langle u_t, \phi \rangle_{(W^{1,p}(B(0)))' \times W^{1,p}(B(0))} - \int_{[u>0]_T} u^n \nabla \Delta u \nabla \phi = 0$$

for $p > \frac{4N}{2N+n(2-N)}$, *arbitrary* $T > 0$, *and for all test functions* ϕ *contained in* $L^2((0,T);W^{1,\infty}(\mathbb{R}^N))$ *such that* $\bigcup_{t\in(0,T)} \mathrm{supp}(\phi(.,t)) \subset B(0)$, *where* $B(0)$ *is an arbitrary ball centered in the origin* $0 \in \mathbb{R}^N$.

(iii) *The solution* u *attains initial data* u_0 *in the sense that*

$$\lim_{t\searrow 0} u(\cdot,t) = u_0(\cdot) \qquad \text{in } L^\beta_{loc}(\mathbb{R}^N) \tag{2.7}$$

for arbitrary $1 \le \beta < \frac{2N}{N-2}$.

(iv) *The solution* u *is an element of* $L^\infty(\mathbb{R}^+;L^\beta(\mathbb{R}^N))$ *for* $1 < \beta < \frac{2N}{N-2}$. *More precisely, a positive constant* $C = C(\beta,N)$ *exists such that the following estimate holds:*

$$\sup_{t\in\mathbb{R}^+} \int_{\mathbb{R}^N} u^\beta(x,t)dx \le C(\beta,N) \left\{ \left(\int_{\mathbb{R}^N} |\nabla u_0|^2 \right)^{1/2} + \int_{\mathbb{R}^N} u_0 \right\}^\beta . \tag{2.8}$$

(v) u *satisfies for arbitrary* $T > 0$ *the basic energy estimate*

$$\begin{aligned} &\int_{\mathbb{R}^N} |\nabla u(\cdot,T)|^2 + C_0^{-1} \int_0^T \int_{\mathbb{R}^N} \left\{ |\nabla u^{\frac{n+2}{6}}|^6 + |\nabla \Delta u^{\frac{n+2}{2}}|^2 \right\} \\ &\qquad \le \int_{\mathbb{R}^N} |\nabla u_0|^2 . \end{aligned} \tag{2.9}$$

Moreover, for arbitrary $R_0 \ge 0$ *and arbitrary* $T > 0$ *the following weighted version of the energy estimate holds with a positive constant* C_1 *that depends only on* n *and* N*:*

$$\begin{aligned} &\int_{\mathbb{R}^N\setminus B_{R_0}(0)} (|x|-R_0)^6 |\nabla u(\cdot,T)|^2 \, dx \\ &\qquad + C_1^{-1} \int_0^T \int_{\mathbb{R}^N\setminus B_{R_0}(0)} (|x|-R_0)^6 \left\{ |\nabla u^{\frac{n+2}{6}}|^6 + |\nabla \Delta u^{\frac{n+2}{2}}|^2 \right\} \\ &\qquad \le \int_{\mathbb{R}^N\setminus B_{R_0}(0)} (|x|-R_0)^6 \cdot |\nabla u_0|^2 \, dx + C_1 \int_0^T \int_{\mathbb{R}^N\setminus B_{R_0}(0)} u^{n+2}. \end{aligned} \tag{2.10}$$

Remark 1. For $n \in (\frac{1}{8}, 2 - \sqrt{8/(8+N)})$, an existence result for the Cauchy problem can be found in Bertsch et al. [8]. However, the solution concept applied in that paper is technically much more involved since third order derivatives are not controlled.

In the course of proof of the result on finite speed of propagation, we need a homogeneous version of Gagliardo–Nirenberg's inequality valid on the complement of balls in $\mathbb{R}^N$. It reads as follows.

LEMMA 2.2. *Let* $1 \le r < \infty$, $0 < q < p$, $m \in \mathbb{N}_+$ *such that*

$$\frac{1}{r} - \frac{m}{N} < \frac{1}{p}.$$

Assume w *to be contained in* $W^{m,r}(\mathbb{R}^N \setminus \overline{B_R(0)}) \cap L^q(\mathbb{R}^N \setminus \overline{B_R(0)})$. *There is a positive constant* $K_1 = K_1(N,m,p,q,r)$ *such that*

$$\|w\|_{p,\mathbb{R}^N\setminus\overline{B_R(0)}} \le K_1 \cdot \|D^m w\|^a_{r;\mathbb{R}^N\setminus\overline{B_R(0)}} \cdot \|w\|^{1-a}_{q,\mathbb{R}^N\setminus\overline{B_R(0)}} . \tag{2.11}$$

Here, $a = (\frac{1}{q} - \frac{1}{p})/(\frac{1}{q} + \frac{m}{N} - \frac{1}{r})$.

Remark 2. With a slight misuse of notation, we write $\|u\|_p$ for $(\int |u|^p)^{1/p}$ also in the case $0 < p < 1$.

Before giving the proof, let us state Gagliardo–Nirenberg's inequality in the following form (see Dal Passo, Giacomelli, and Shishkov [11]).

LEMMA 2.3. *Let* $1 \le r \le \infty$, $0 < q < p$, $m \in \mathbb{N}_+$ *such that*

$$\frac{1}{r} - \frac{m}{N} < \frac{1}{p}.$$

If $\Omega \subset \mathbb{R}^N$ *is bounded with piecewise smooth boundary, then positive constants* c_1 *and* c_2 *depending only on* $\Omega, r, p, m,$ *and* q *exist such that for any* $u \in L^q(\Omega)$ *satisfying* $D^m u \in L^r(\Omega)$, *the following inequality holds:*

$$\|u\|_p \le c_1 \|D^m u\|_r^a \|u\|_q^{1-a} + c_2 \|u\|_q, \tag{2.12}$$

where $a = (\frac{1}{q} - \frac{1}{p})/(\frac{1}{q} + \frac{m}{N} - \frac{1}{r})$.

Especially, if Ω *is an infinite cone, i.e., for given points* $x_0, y_0 \in \mathbb{R}^N$, $x_0 \notin B_1(y_0)$ *a set*

$$C_{x_0,y_0} := \left\{ z \in \mathbb{R}^N | z = x_0 + \lambda(y - x_0), \quad y \in B_1(y_0), \quad \lambda > 0 \right\},$$

then (2.12) *holds with constants* $c_1 = c(\|x_0 - y_0\|, r, p, m, q)$ *and* $c_2 = 0$.

Proof of Lemma 2.2. Let us prove the result first for the special case $\Omega = \mathbb{R}^N \setminus \overline{B_1(0)}$. To this purpose, we write

$$\Omega = \Omega_+ \cup \Omega_-,$$

where Ω_+ and Ω_- are open sets which are $W^{m,\infty}$-diffeomorphic to the half-space $\mathbb{R}^N_+ := \{x \in \mathbb{R}^N | x_N > 0\}$. In the case $m = 1$, for instance, we may choose Ω_+ and Ω_- as the complements in $\mathbb{R}^N$ of the closed sets

$$A_+ := \overline{B_1(0)} \cup \left\{ x \in \mathbb{R}^N | x_N \ge 0 \right\}$$

and

$$A_- := \overline{B_1(0)} \cup \left\{ x \in \mathbb{R}^N | x_N \le 0 \right\},$$

respectively. For $m > 1$, an appropriate smoothing procedure has to be applied.

By virtue of Lemma 2.3 and a straightforward transformation argument, a Gagliardo–Nirenberg inequality in the spirit of (2.11) holds both on Ω_+ and on Ω_-.

This is sufficient to prove (2.11) for $\Omega = \mathbb{R}^N \setminus \overline{B_1(0)}$. Indeed,

$$\begin{aligned}
\int_\Omega |w|^p &\le \int_{\Omega_+} |w|^p + \int_{\Omega_-} |w|^p \\
&\le C\Bigg\{ \left(\int_{\Omega_+} |D^m w|^r \right)^{\frac{ap}{r}} \cdot \left(\int_{\Omega_+} |w|^q \right)^{\frac{p}{q}(1-a)} \\
&\qquad + \left(\int_{\Omega_-} |D^m w|^r \right)^{\frac{ap}{r}} \cdot \left(\int_{\Omega_-} |w|^q \right)^{\frac{p}{q}(1-a)} \Bigg\} \\
&\le C \left(\int_{\Omega_+} |D^m w|^r + \int_{\Omega_-} |D^m w|^r \right)^{\frac{ap}{r}} \left(\int_{\Omega_+} |w|^q + \int_{\Omega_-} |w|^q \right)^{\frac{p}{q}(1-a)} \\
&\le K_1 \left(\int_\Omega |D^m w|^r \right)^{\frac{ap}{r}} \left(\int_\Omega |w|^q \right)^{\frac{p}{q}(1-a)} .
\end{aligned}$$

In the second step of this estimate, we used the assumption

$$\frac{1}{r} - \frac{m}{N} < \frac{1}{p} < \frac{1}{q}$$

together with the calculus inequality

$$\sum_{i=1}^k a_i^\alpha b_i^\beta \le \left(\sum_{i=1}^k a_i \right)^\alpha \left(\sum_{i=1}^k b_i \right)^\beta ,$$

which holds for numbers α, β and a_i, b_i, $i = 1, \dots, k$, that satisfy

$$a_i, b_i \ge 0, \quad \alpha, \beta > 0, \quad \text{and} \quad \alpha + \beta \ge 1.$$

For a proof, see, for instance, Dal Passo, Giacomelli, and Shishkov [11]. Finally, (2.11) follows for arbitrary $R > 0$ by a straightforward scaling argument. □

3. An application of Hardy's inequality. In this section we will prove a Hardy-type estimate on $\mathbb{R}^N \setminus B_R(0)$, which will be combined with the weighted energy-type estimate (2.10) to yield a key ingredient for the proof of the main result of the paper. The Hardy-type estimate reads as follows.

LEMMA 3.1. *Assume that $w(r) := r^4$ and $v(r) := r^6$ and that $u \in H^1_{loc}(\mathbb{R}^N) \cap C(\mathbb{R}^N)$ satisfies the inequalities*

$$\int_{\mathbb{R}^N} v(|x|) \cdot |\nabla u|^2 \, dx < \infty, \tag{3.1}$$

$$\int_{\partial B_R(0)} u^2 \, d\mathcal{H}^{N-1} \le C_1 \cdot R^{-1} . \tag{3.2}$$

Then a positive constant C_2 exists which is independent of $R > 0$ such that

$$\int_{\mathbb{R}^N \setminus B_R(0)} w(\operatorname{dist}(x, B_R(0))) \cdot u^2 \, dx \le C_2 \int_{\mathbb{R}^N \setminus B_R(0)} v(\operatorname{dist}(x, B_R(0))) |\nabla u|^2 \, dx . \tag{3.3}$$

Let us first recall the following version of Hardy's inequality in one space dimension (see [23], [24], and the monograph [28]).

LEMMA 3.2. *Let a be a real number and assume the weight functions v, w to be nonnegative and measurable on (a, ∞). Consider for $x \in (a, \infty)$ the quantity*

$$F_{Har}(x) := \int_a^x w(s)ds \cdot \int_x^\infty v^{-1}(s)ds.$$

If $\sup_{a<x<\infty} F_{Har}(x) < \infty$, then there exists a positive constant $C = C(a, v, w)$ such that

$$\int_a^\infty w(s) \cdot u^2(s)ds \le C \int_a^\infty v(s) \cdot u_x^2(s)ds \tag{3.4}$$

for all

$$u \in AC_R(a, \infty) := \{u \in W^{1,1}_{loc}(a, \infty) : \lim_{x \nearrow \infty} u(x) = 0\}.$$

Proof of Lemma 3.1. The strategy is to apply the one-dimensional Hardy inequality (3.4) to appropriate integrals over spheres of radius r. Therefore, let us switch to polar coordinates and prove an estimate on the derivative of the L^2-norm of u over such spheres. First we need some notation.

$$\begin{aligned}\int_{\mathbb{R}^N \setminus B_R(0)} w(\operatorname{dist}(x, B_R(0))u^2\, dx &= \int_R^\infty \int_{S^{N-1}} w(r-R)u^2 r^{N-1}\, dS^{N-1}\, dr \\ &= \int_R^\infty w(r-R)U^2(r)r^{N-1}\, dr,\end{aligned} \tag{3.5}$$

where we defined

$$U(r) := \left(\int_{S^{N-1}} u^2(r, \theta)\, dS^{N-1}\right)^{1/2}.$$

Note that we use "u" to denote the function u in both polar and Euclidean coordinates. Here, "θ" is an abbreviation of the angular coordinates and "dS^{N-1}" stands for the surface element on the unit sphere. We have in particular that

$$\frac{\partial}{\partial r}U(r) = \left(\int_{S^{N-1}} u^2(r, \theta)\, dS^{N-1}\right)^{-1/2} \cdot \underbrace{\left(\int_{S^{N-1}} u(r, \theta) \cdot u_r(r, \theta) \cdot dS^{N-1}\right)}_{\le \left(\int_{S^{N-1}} u^2\, dS^{N-1}\right)^{1/2}\left(\int_{S^{N-1}} u_r^2\, dS^{N-1}\right)^{1/2}}, \tag{3.6}$$

which implies that

$$\left|\frac{\partial}{\partial r}U(r)\right|^2 \le \int_{S^{N-1}} u_r^2\, dS^{N-1}. \tag{3.7}$$

On the other hand, (3.2) entails the decay estimate

$$U(r) \le C \cdot r^{-\frac{N}{2}}$$

for $r > 0$. Altogether,

$$U(r) \in AC_R(R, \infty).$$

Now assuming that we may apply Hardy's inequality for the weight functions

$$\tilde{w}(r) := (r-R)^4 \cdot r^{N-1}$$

and

$$\tilde{v}(r) := (r-R)^6 \cdot r^{N-1},$$

we obtain the following result by virtue of Lemma 3.2:

$$\begin{aligned}\int_{\mathbb{R}^N \setminus B_R(0)} w(\mathrm{dist}(x, B_R(0)))u(x)^2\, dx &= \int_R^\infty \tilde{w}(r) \cdot U^2(r) dr \\ &\le C \int_R^\infty \tilde{v}(r) \cdot U_r^2(r) dr \\ &\quad \text{(by (3.7))} \\ &\le \int_{\mathbb{R}^N \setminus B_R(0)} v(\mathrm{dist}(x, B_R(0)))|\nabla u|^2\, dx\,.\end{aligned} \tag{3.8}$$

It remains to verify that we were indeed allowed to use Hardy's inequality. This means we have to convince ourselves that

$$\sup_{R<x<\infty} \left\{ \int_R^x \tilde{w}(s) ds \cdot \int_x^\infty \tilde{v}^{-1}(s) ds \right\} < \infty\,.$$

Indeed, we find for arbitrary $x \in (R, \infty)$ that

$$\begin{aligned}&\int_R^x (r-R)^4 \cdot r^{N-1}\, dr \cdot \int_x^\infty (r-R)^{-6} r^{1-N}\, dr \\ &\le x^{N-1} \cdot \int_R^x (r-R)^4\, dr \cdot x^{1-N} \int_x^\infty (r-R)^{-6}\, dr \\ &\le \int_R^x (r-R)^4\, dr \cdot \int_x^\infty (r-R)^{-6} \\ &= \frac{1}{25}\,,\end{aligned}$$

and the lemma is proved. □

Lemma 3.1 can be combined with Theorem 2.1 to establish the following estimate on solutions to the Cauchy problem.

Lemma 3.3. *Let u be a solution to the Cauchy problem as constructed in Theorem 2.1. Then a positive constant C_2 which is independent of $R_0 \ge 0$ exists such that*

$$\begin{aligned}&\int_{\mathbb{R}^N \setminus B_{R_0}(0)} (|x| - R_0)^4 u(\cdot, T)^2\, dx \\ &\quad + C_2^{-1} \int_0^T \int_{\mathbb{R}^N \setminus B_{R_0}(0)} (|x| - R_0)^6 \left\{ |\nabla u^{\frac{n+2}{6}}|^6 + |\nabla \Delta u^{\frac{n+2}{2}}|^2 \right\} \\ &\quad \le C_2 \left\{ \int_{\mathbb{R}^N \setminus B_{R_0}(0)} (|x| - R_0)^6 \cdot |\nabla u_0|^2\, dx + \int_0^T \int_{\mathbb{R}^N \setminus B_{R_0}(0)} u^{n+2} \right\}.\end{aligned} \tag{3.9}$$

Proof. The proof will be an immediate consequence of Lemma 3.1 and the weighted energy estimate (2.10), provided we can show a decay estimate like (3.2) for $\int_{\partial B_R(0)} u^2 d\mathcal{H}^{N-1}$.

Note that the weighted energy estimate (2.10) and the global energy estimate

$$\int_{\mathbb{R}^N} |\nabla u(\cdot,T)|^2 + C_1 \int_0^T \int_{\mathbb{R}^N} \left\{ |\nabla u^{\frac{n+2}{6}}|^6 + |\nabla \Delta u^{\frac{n+2}{2}}|^2 \right\} \le \int_{\mathbb{R}^N} |\nabla u_0|^2$$

imply that

$$\sup_{0<t<T} \int_{\mathbb{R}^N} |x|^2 |\nabla u(x,t)|^2 < \infty. \tag{3.10}$$

On the other hand, (2.8) entails that

$$\sup_{0<t<T} \int_{\mathbb{R}^N} u^2(x,t) dx < \infty.$$

Then the result follows by application of the following lemma. ∎

LEMMA 3.4. *Assume that a function $v \in H^1(\mathbb{R}^N)$ satisfies*

$$\int_{\mathbb{R}^N} x^2 |\nabla v|^2 \, dx + \int_{\mathbb{R}^N} v^2 \, dx < \infty. \tag{3.11}$$

Then there is a positive constant C_3 such that

$$\int_{\partial B_R(0)} v^2 d\mathcal{H}^{N-1} \le C_2 \cdot R^{-1} \tag{3.12}$$

for arbitrary $R > 0$.

Proof. Let us denote the transformation of v to polar coordinates by $\hat{v}$. Consider the quantity

$$r \cdot \|\hat{v}\|^2_{2,\partial B_r(0)} := r \cdot \int_{S^{N-1}} \hat{v}^2(r,\theta) \cdot r^{N-1} \, dS^{N-1} \,. \tag{3.13}$$

Differentiation with respect to r gives

$$\frac{\partial}{\partial r} \left(r \|\hat{v}\|^2_{2,\partial B_r(0)} \right) = 2r \int_{S^{N-1}} \hat{v} \hat{v}_r r^{N-1} \, dS^{N-1} + N \int_{S^{N-1}} \hat{v}^2 r^{N-1} \, dS^{N-1} \,.$$

Integrating this identity over the interval $(0,R)$ with respect to r implies that

$$\begin{aligned}
R \int_{S^{N-1}} & \hat{v}^2 R^{N-1} \, dS^{N-1} \\
&= 2 \int_0^R \int_{S^{N-1}} r \hat{v} \hat{v}_r r^{N-1} \, dS^{N-1} \, dr + N \int_0^R \int_{S^{N-1}} \hat{v}^2 r^{N-1} \, dS^{N-1} \, dr \\
&\le C \left(\int_0^R \int_{S^{N-1}} r^2 \hat{v}_r^2 r^{N-1} \, dS^{N-1} \, dr + \int_0^R \int_{S^{N-1}} \hat{v}^2 r^{N-1} \, dS^{N-1} \, dr \right) \\
&\le C \left(\int_{B_R} |x|^2 |\nabla v|^2 \, dx + \int_{B_R} v^2 \, dx \right) .
\end{aligned}$$

This proves the assertion of Lemma 3.4. ∎

4. The main result. This section is devoted to the proof of Theorem 4.1.

THEOREM 4.1. *Let $n \in (2 - \sqrt{8/(8+N)}, 3)$, $N < 4$, and assume initial data $u_0 \in H^1(\mathbb{R}^N)$ to be nonnegative with compact support in the sense that $u_0(x) = 0$ almost everywhere on $\mathbb{R}^N \setminus B_{R_0}(0)$ for a positive number R_0. Let u be the strong solution to the Cauchy problem constructed in* Theorem 2.1. *Then u has finite speed of propagation in the sense of Definition* 1.1. *More precisely,* $\mathrm{supp}(u(\cdot,t)) \subset B_{R(t)}(0)$, *where*

$$R(t) = R_0 + C \cdot t^{\frac{1}{\alpha}} \left(\int_0^t \left(\int_{\mathbb{R}^N \setminus B_{R_0}} u^2 \right)^{\frac{n+2}{2}} \right)^{\frac{n}{2\alpha}} \tag{4.1}$$

with a positive constant $C = C(n,N)$ and $\alpha = \frac{(8+Nn)(n+2)}{4}$.

Remark 3. 1. The notion of *finite speed of propagation* formulated in Definition 1.1 is still a rather weak one. However, it is possible to replace balls by general convex sets having sufficiently smooth boundary. With—sometimes rather tedious—technical changes, the results to be proved in this section continue to hold. For further improvements, e.g., the treatment of initial data with nonconvex support, refined versions of Hardy's inequality will be necessary. In the forthcoming paper [17], we will prove a Hardy-type inequality valid on infinite cones. On the basis of the finite speed of propagation result established in the present paper, new weighted energy estimates will be formulated for which the spatial support will be given by an infinite cone. These estimates will be the key ingredient for proving local results both on finite speed of propagation and on the occurrence of a waiting time phenomenon wherever the support of initial data locally satisfies an exterior cone condition.

2. Note that (4.1) can be combined with the global energy estimate (2.9) to yield the estimate

$$R(t) \le R_0 + \hat{C} \cdot t^{\frac{2}{8+nN}}$$

with a constant $\hat{C}$ depending on $n, N, \|\nabla u_0\|_2$, and the initial mass. However, the exponent $\gamma = \frac{2}{8+nN}$ is not optimal as a comparison with self-similar solutions reveals (see Ferreira and Bernis [13]). Nevertheless, the merely qualitative result presented here is the starting point for providing optimal quantitative estimates on the diameter of $\mathrm{supp}(u(\cdot,t))$. This is the subject of the forthcoming paper [22].

Proof of Theorem 4.1. The starting point is the estimate (3.9), which can be simplified for arbitrary $R \ge R_0$ and $T > 0$ in the following way:

$$\begin{aligned} \sup_{t\in(0,T)} \int_{\mathbb{R}^N \setminus B_R} (|x| - R)^4 u(\cdot,t)^2 + C_2^{-1} \int_0^T \int_{\mathbb{R}^N \setminus B_R} (|x| - R)^6 |\nabla u^{\frac{n+2}{6}}|^6 \\ \le C_2 \int_0^T \int_{\mathbb{R}^N \setminus B_R} u^{n+2}. \end{aligned} \tag{4.2}$$

Consider positive numbers $\varrho > R$. Obviously, $|x| - R > \varrho - R$ on $\mathbb{R}^N \setminus B_\varrho$. This implies that

$$\begin{aligned} \sup_{t\in(0,T)} \int_{\mathbb{R}^N \setminus B_\varrho} u(\cdot,t)^2 + (\varrho - R)^2 \int_0^T \int_{\mathbb{R}^N \setminus B_\varrho} |\nabla u^{\frac{n+2}{6}}|^6 \\ \le \frac{C}{(\varrho - R)^4} \int_0^T \int_{\mathbb{R}^N \setminus B_R} u^{n+2} \end{aligned} \tag{4.3}$$

for $\varrho > R \geq R_0$.

By virtue of Lemma 2.2, the term on the right-hand side can be estimated as follows:

(4.4)
$$\int_0^T \int_{\mathbb{R}^N \setminus B_R} u^{n+2} \leq K_1 \left(\int_0^T \int_{\mathbb{R}^N \setminus B_R} |\nabla u^{\frac{n+2}{6}}|^6 \right)^{\frac{nN}{nN+12}} \left(\int_0^T \left(\int_{\mathbb{R}^N \setminus B_R} u^2 \right)^{\frac{n+2}{2}} \right)^{\frac{12}{nN+12}}.$$

Young's inequality yields

(4.5)
$$\begin{aligned} \frac{1}{(\varrho - R)^4} \int_0^T \int_{\mathbb{R}^N \setminus B_R} u^{n+2} &\leq (\varrho - R)^{\frac{2nN}{nN+12}} K_1 \left(\int_0^T \int_{\mathbb{R}^N \setminus B_R} |\nabla u^{\frac{n+2}{6}}|^6 \right)^{\frac{nN}{nN+12}} \\ &\quad \cdot (\varrho - R)^{-4-\frac{2nN}{nN+12}} \left(\int_0^T \left(\int_{\mathbb{R}^N \setminus B_R} u^2 \right)^{\frac{n+2}{2}} \right)^{\frac{12}{nN+12}} \\ &\leq \varepsilon (\varrho - R)^2 \int_0^T \int_{\mathbb{R}^N \setminus B_R} |\nabla u^{\frac{n+2}{6}}|^6 \\ &\quad + C_\varepsilon (\varrho - R)^{-(4+\frac{nN}{2})} \int_0^T \left(\int_{\mathbb{R}^N \setminus B_R} u^2 \right)^{\frac{n+2}{2}}. \end{aligned}$$

Putting everything together gives

(4.6)
$$\begin{aligned} &\sup_{t \in (0,T)} \int_{\mathbb{R}^N \setminus B_\varrho} u^2(\cdot, t) + (\varrho - R)^2 \int_0^T \int_{\mathbb{R}^N \setminus B_\varrho} |\nabla u^{\frac{n+2}{6}}|^6 \\ &\qquad \leq \varepsilon (\varrho - R)^2 \int_0^T \int_{\mathbb{R}^N \setminus B_R} |\nabla u^{\frac{n+2}{6}}|^6 + \frac{C_\varepsilon}{(\varrho - R)^{4+\frac{nN}{2}}} \int_0^T \left(\int_{\mathbb{R}^N \setminus B_R} u^2 \right)^{\frac{n+2}{2}} \end{aligned}$$

for all $\varrho > R > R_0$.

Introducing the quantities

$$V(\varrho) := \sup_{t \in (0,T)} \int_{\mathbb{R}^N \setminus B_\varrho} u^2, \qquad U(\varrho) := \int_0^T \int_{\mathbb{R}^N \setminus B_\varrho} |\nabla u^{\frac{n+2}{6}}|^6,$$
$$F_\varepsilon(\varrho, R) := \frac{C_\varepsilon}{(\varrho - R)^{4+\frac{nN}{2}}} \int_0^T \left(\int_{\mathbb{R}^N \setminus B_R} u^2 \right)^{\frac{n+2}{2}},$$

(4.6) can be written as follows:

$$V(\varrho) + (\varrho - R)^2 U(\varrho) \leq \varepsilon (\varrho - R)^2 U(R) + F_\varepsilon(\varrho, R)$$

for all $\varepsilon > 0$ and all $\varrho > R \geq R_0$. An application of the subsequent iteration result Lemma 4.2 shows that

$$V(\varrho) + \frac{(\varrho - R)^2}{4} U(\varrho) \leq K_\varepsilon F_\varepsilon(\varrho, R)$$

for sufficiently small but fixed $\varepsilon > 0$. Therefore,

$$\sup_{t\in(0,T)} \int_{\mathbb{R}^N\setminus B_\varrho} u^2(\cdot,t) + (\varrho - R)^2 \int_0^T \int_{\mathbb{R}^N\setminus B_\varrho} |\nabla u^{\frac{n+2}{6}}|^6 \leq \frac{K_\varepsilon}{(\varrho-R)^{4+\frac{nN}{2}}} \int_0^T \left(\int_{\mathbb{R}^N\setminus B_R} u^2 \right)^{\frac{n+2}{2}}. \tag{4.7}$$

Using the estimate

$$\int_0^T \left(\int_{\mathbb{R}^N\setminus B_\varrho} u^2 \right)^{\frac{n+2}{2}} \leq T \cdot \sup_{t\in(0,T)} \left(\int_{\mathbb{R}^N\setminus B_\varrho} u^2 \right)^{\frac{n+2}{2}},$$

we find that

$$\int_0^T \left(\int_{\mathbb{R}^N\setminus B_\varrho} u^2 \right)^{\frac{n+2}{2}} \leq \frac{CT}{(\varrho-R)^{(4+\frac{nN}{2})\frac{n+2}{2}}} \cdot \left(\int_0^T \left(\int_{\mathbb{R}^N\setminus B_R} u^2 \right)^{\frac{n+2}{2}} \right)^{\frac{n+2}{2}}. \tag{4.8}$$

Introducing

$$G(\varrho) := \int_0^T \left(\int_{\mathbb{R}^N\setminus B_\varrho} u^2 \right)^{\frac{n+2}{2}},$$

$$\alpha := \left(4 + \frac{nN}{2}\right)\frac{n+2}{2}, \qquad \beta := \frac{n+2}{2} > 1,$$

(4.8) can be rewritten in the form

$$G(\varrho) \leq \frac{CT}{(\varrho-R)^\alpha} \cdot G(R)^\beta$$

for all $\varrho > R \geq R_0$. An application of Stampacchia's iteration lemma (Lemma 4.3) shows that $G(\varrho) = 0$, provided

$$(\varrho - R_0)^\alpha \geq C \cdot T \cdot \left(\int_0^T \left(\int_{\mathbb{R}^N\setminus B_{R_0}} u^2 \right)^{\frac{n+2}{2}} \right)^{\frac{n}{2}}.$$

Hence, we obtain for $R(T)$ the estimate given in equation (4.1). Recalling in addition that $u \in L^\infty(\mathbb{R}^+; L^\beta(\mathbb{R}^N))$ for all $1 < \beta < \frac{2N}{N-2}$, we have proved finite speed of propagation. □

We used the following iteration lemma, which is a slight modification of an argument presented in the proof of Theorem 6.1 of [10] (see also [25]). For the reader's convenience, we sketch the proof.

LEMMA 4.2. *Assume that*

$$V(\varrho') + (\varrho' - R')^2 U(\varrho') \leq \varepsilon(\varrho' - R')^2 U(R') + F_\varepsilon(\varrho', R') \tag{4.9}$$

for $\varepsilon > 0$ sufficiently small and $0 \leq R_0 \leq R < R' < \varrho' < \varrho$. Then there exists a positive constant K_ε such that

$$V(\varrho) + \frac{(\varrho - R)^2}{4} U(\varrho) \leq K_\varepsilon F_\varepsilon(\varrho, R). \tag{4.10}$$

Proof. Introduce for $k \in \mathbb{N}$ points ϱ_k, R_k such that

$$\varrho_k = R + \frac{(\varrho - R)}{2^{k-1}}, \qquad R_k := R + \frac{(\varrho - R)}{2^k}, \tag{4.11}$$

i.e.,

$$R_k = \varrho_{k+1} \qquad \text{and} \qquad \varrho_k - R_k = \frac{(\varrho - R)}{2^k}. \tag{4.12}$$

Along the lines of the corresponding result in [10], we may prove that

$$V(\varrho) + \frac{(\varrho - R)^2}{4} U(\varrho) \le \frac{\varepsilon^M}{4} (\varrho - R)^2 U(R_M) + \sum_{k=1}^{M} (4\varepsilon)^{k-1} F_\varepsilon(\varrho_k, R_k). \tag{4.13}$$

Now estimating

$$F_\varepsilon(\varrho_k, R_k) \le 2^{k(4+\frac{nN}{2})} \frac{C_\varepsilon}{(\varrho - R)^{4+\frac{nN}{2}}} \int_0^T \left(\int_{\mathbb{R}^N \setminus B_R} u^2 \right)^{\frac{n+2}{2}},$$

it becomes evident that for ε sufficiently small, the right-hand side of (4.13) is convergent, which proves the lemma. ∎

For the sake for completeness, we state the following.

LEMMA 4.3 (Stampacchia's iteration lemma). *Assume that a given nonnegative, nonincreasing function $G : (0, \varrho_0) \to \mathbb{R}$ satisfies*

$$G(\xi) \le \frac{c_0}{(\xi - \eta)^\alpha} G(\eta)^\beta$$

for $0 \le \eta < \xi \le \varrho_0$ and positive numbers c_0, α, β with $\beta > 1$. Assume further that

$$\varrho_0^\alpha \ge 2^{\frac{\alpha\beta}{\beta-1}} \cdot c_0 \cdot G(0)^{\beta-1}.$$

Then G has a root in ϱ_0.

REFERENCES

[1] E. BERETTA, M. BERTSCH, AND R. DAL PASSO, *Nonnegative solutions of a fourth order nonlinear degenerate parabolic equation,* Arch. Ration. Mech. Anal., 129 (1995), pp. 175–200.

[2] F. BERNIS, *Viscous flows, fourth order nonlinear degenerate parabolic equations and singular elliptic problems,* in Free Boundary Problems: Theory and Applications, J.I. Diaz, M.A. Herrero, A. Linan, and J.L. Vazquez, eds., Pitman Research Notes in Math. 323, Longman Scientific and Technical, Harlow, UK, 1995, pp. 40–56.

[3] F. BERNIS, *Finite speed of propagation and continuity of the interface for thin viscous flows,* Adv. Differential Equations, 1 (1996), pp. 337–368.

[4] F. BERNIS, *Finite speed of propagation for thin viscous flows when $2 \le n < 3$,* C.R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 1169–1174.

[5] F. BERNIS AND A. FRIEDMAN, *Higher order nonlinear degenerate parabolic equations,* J. Differential Equations, 83 (1990), pp. 179–206.

[6] F. BERNIS, L.A. PELETIER, AND S.M. WILLIAMS, *Source-type solutions of a fourth order nonlinear degenerate parabolic equation,* Nonlinear Anal., 18 (1992), pp. 217–234.

[7] A.L. BERTOZZI AND M. PUGH, *The lubrication approximation for thin viscous films: Regularity and long time behaviour of weak solutions,* Comm. Pure Appl. Math., 49 (1996), pp. 85–123.

[8] M. BERTSCH, R. DAL PASSO, H. GARCKE, AND G. GRÜN, *The thin viscous flow equation in higher space dimensions,* Adv. Differential Equations, 3 (1998), pp. 417–440.

[9] R. Dal Passo, H. Garcke, and G. Grün, *On a fourth-order degenerate parabolic equation: Global entropy estimates, existence, and qualitative behavior of solutions,* SIAM J. Math. Anal., 29 (1998), pp. 321–342.

[10] R. Dal Passo, L. Giacomelli, and G. Grün, *A waiting time phenomenon for thin film equations,* Ann. Scuola Norm. Sup. Pisa, 30 (2001), pp. 437–463.

[11] R. Dal Passo, L. Giacomelli, and A. Shishkov, *The thin film equation with nonlinear diffusion,* Comm. Partial Differential Equations, 26 (2001), pp. 1509–1557.

[12] E.B. Dussan and S. Davis, *On the motion of a fluid-fluid interface along a solid surface,* J. Fluid Mech., 65 (1974), pp. 71–95.

[13] R. Ferreira and F. Bernis, *Source-type solutions to thin-film equations in higher space dimensions,* European J. Appl. Math., 8 (1997), pp. 507–524.

[14] L. Giacomelli and F. Otto, *Droplet spreading: Intermediate scaling law by pde methods,* Comm. Pure Appl. Math., 55 (2002), pp. 217–254.

[15] H.P. Greenspan, *On the motion of a small viscous droplet that wets a surface,* J. Fluid Mech, 84 (1978), pp. 125–143.

[16] H.P. Greenspan and B.M. McKay, *On the wetting of a surface by a very viscous fluid,* Stud. Appl. Math., 64 (1981), pp. 95–112.

[17] G. Grün, *Droplet spreading under weak slippage: The waiting time phenomenon,* submitted for publication.

[18] G. Grün, *On the convergence of entropy consistent schemes for lubrication type equations in multiple space dimensions,* Math. Comp., to appear.

[19] G. Grün, *Degenerate parabolic equations of fourth order and a plasticity model with nonlocal hardening,* Z. Anal. Anwendungen, 14 (1995), pp. 541–573.

[20] G. Grün, *On Bernis' interpolation inequalities in multiple space dimensions,* Z. Anal. Anwendungen, 20 (2001), pp. 987–998.

[21] G. Grün, *On Free Boundary Problems Arising in Thin Film Flow,* Habilitation thesis, University of Bonn, Bonn, Germany, 2001.

[22] G. Grün, *Droplet spreading under weak slippage: The optimal asymptotic propagation rate in the multi-dimensional case,* Interfaces Free Bound., 4 (2002), pp. 309–323.

[23] G.H. Hardy, *Note on a theorem of Hilbert,* Math. Z., 6 (1920), pp. 314–317.

[24] G.H. Hardy, J.E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1934.

[25] J. Hulshof and A. Shishkov, *The thin film equation with $2 \le n < 3$: Finite speed of propagation in terms of the L^1-norm,* Adv. Differential Equations, 3 (1998), pp. 625–642.

[26] A.V. Ivanov, *Existence and uniqueness of a regular solution of the Cauchy-Dirichlet problem for doubly nonlinear parabolic equations,* Z. Anal. Anwendungen, 14 (1995), pp. 751–777.

[27] P. Neogi and C.A. Miller, *Spreading kinetics of a drop on a rough solid surface,* J. Colloid Interface Sci., 92 (1983), pp. 338–349.

[28] B. Opic and A. Kufner, *Hardy-Type Inequalities*, Pitman Research Notes in Math. 219, Longman Scientific and Technical, Harlow, UK, 1990.

[29] A. Oron, S.H. Davis, and S.G. Bankoff, *Long-scale evolution of thin liquid films,* Rev. Modern Phys., 69 (1997), pp. 932–977.

[30] J.L. Vazquez, *An introduction to the mathematical theory of the porous medium equation,* in Shape Optimization and Free Boundaries, M.C. Delfour and G. Sabidussi, eds., Kluwer Academic, Dordrecht, The Netherlands, 1992, pp. 347–389.

SIAM J. MATH. ANAL.
Vol. 34, No. 5, pp. 1007–1039

ROBUST PERMANENCE FOR ECOLOGICAL DIFFERENTIAL EQUATIONS, MINIMAX, AND DISCRETIZATIONS*

BARNABAS M. GARAY† AND JOSEF HOFBAUER‡

Abstract. We present a sufficient condition for robust permanence of ecological (or Kolmogorov) differential equations based on average Liapunov functions. Via the minimax theorem we rederive Schreiber's sufficient condition [S. Schreiber, *J. Differential Equations*, 162 (2000), pp. 400–426] in terms of Liapunov exponents and give various generalizations. Then we study robustness of permanence criteria against discretizations with fixed and variable stepsizes. Applications to mathematical ecology and evolutionary games are given.

Key words. population dynamics, average Liapunov functions, robust permanence, discretizations of Kolmogorov type, invariant measures, Liapunov exponent, minimax, Conley index, evolutionary games

AMS subject classifications. Primary, 37B25, 65L05; Secondary, 37B30, 92D25, 90C47, 91A22

PII. S0036141001392815

1. Introduction.

The concept of permanence (also known as uniform persistence) emerged in the late seventies as the appropriate mathematical description of coexistence in deterministic models of interacting species, replacing the previously used, but far too restrictive, global asymptotic stability of an equilibrium. It simply requires that the boundary of the state space, or the set of all extinction states, be a repeller for the dynamics of the ecological system.

In the late eighties it was realized [16], [28] that the proper framework for permanence (for the boundary as a whole) in topological dynamics was already developed by Zubov, Ura, Kimura, and others (see historical remarks in section 2), while Conley's Morse decompositions allow a finer description. New ideas of Schreiber [57] in a C^r setting are the use of invariant measures and ergodic theory, in particular smooth ergodic theory, and lead to characterizations of a robust form of permanence, meaning that nearby systems are still permanent.

In the present paper we derive sufficient conditions for robust permanence along a more classical approach using topological dynamics, in particular "good" average Liapunov functions (GALF), the Zubov–Ura–Kimura theorem, and Morse decompositions. Our key result is to relate the standard average Liapunov functions $P(x) = \prod x_i^{p_i}$ via the minimax theorem to invariant measures. This allows us to rederive and strengthen Schreiber's [57] sufficient conditions stated in terms of "unsaturated" invariant measures. (Our paper does not concern Schreiber's necessary conditions for robust permanence, based on the deep theory of measurable stable manifolds of Pesin.) Our approach leads to sharper robustness results: First, we allow C^0-perturbations; second, we prove uniform separation of the dual attractor from the repelling boundary. Similar sharper results were recently and independently obtained also by Hirsch, Smith, and Zhao [26] by refining the invariant measure ap-

*Received by the editors July 24, 2001; accepted for publication (in revised form) October 31, 2002; published electronically April 15, 2003.
http://www.siam.org/journals/sima/34-5/39281.html
†Department of Mathematics, University of Technology, Budapest, Hungary (garay@math.bme.hu).
‡Department of Mathematics, University of Vienna, Austria (Josef.Hofbauer@univie.ac.at).

proach of [57]. However, our approach, based on GALF, is suitable to derive explicit estimates; see Remark 2.7. It also leads to exponential repulsivity; see section 3, where we also shed light on the relation between GALF and Liapunov functions.

We work out the details for dynamics on the probability simplex and indicate that similar results hold for ecological (or Kolmogorov) systems in $\mathbb{R}^n_+$ (as in [57]) and also for systems with a compact codimension 1 invariant manifold; see section 11.1.

In the second part of the paper we transfer the previous results to discrete-time systems and then turn our attention to discretization problems. We show that natural discretizations of ecological differential equations respect the invariance of boundary faces. For such discretizations of Kolmogorov type we prove robustness of permanence and discuss also Kloeden–Schmalfuss pullback attractor–repeller pairs with free step-size sequences.

In section 10 we give a short review (including open problems) of the literature on index theorems ensuring that isolated invariant sets on the boundary actually repel trajectories into the interior.

In the final section 11 we illustrate the theory with a number of applications to ecological and game theoretic models, such as Lotka–Volterra equations, replicator equations, and imitation dynamics, as well as their discretized versions. Other applications concern invasion of an ecological system by a new species and explicit characterizations of totally permanent systems which are robustly permanent together with all their subsystems.

We use the terminology of standard textbooks like Conley [12] and Nemytzkii and Stepanov [55] without any further notice. In particular, we use the terms attractor and repeller as in [12]. Index theory and ergodic theory of dynamical systems, used in this paper, are contained in these two monographs. We recommend also [2], [51], and [62].

Notation. The nonnegative orthant in $\mathbb{R}^n$ is denoted by $\mathbb{R}^n_+$ and the positive orthant by $\operatorname{int}\mathbb{R}^n_+$. The boundary, closure, and interior of a subset $S \subset X$ are denoted by ∂S, $\operatorname{cl}(S)$, and $\operatorname{int}(S)$. $\mathcal{B}[A,\varepsilon] = \{x : d(x,A) \le \varepsilon\}$ and $\mathcal{B}(A,\varepsilon) = \{x : d(x,A) < \varepsilon\}$ denote the closed and open ε-neighborhood of a set A.

Capital Greek letters Φ, Ψ denote continuous-time dynamical systems. Discretizations are denoted by the respective lowercase Greek letters. In dynamical concepts like $\gamma^+_\Phi(x)$, $\mathcal{A}_\Psi$, $\omega_{\varphi(h,\cdot)}(x)$, etc., the subscripts refer to the corresponding continuous-time or discrete-time dynamical systems.

2. Robust permanence. We consider an autonomous differential equation of Kolmogorov type,

$$\dot{x}_i = x_i f_i(x), \quad x \in X, \tag{1}$$

where X is the probability simplex $\{x \in \mathbb{R}^n : x_i \ge 0, \sum_i x_i = 1\}$ and $f : X \to \mathbb{R}^n$ is a continuous function satisfying $\sum_i x_i f_i(x) = 0$ for each $x \in X$. The standard interpretation in biology is that x_i represents the proportion of the ith species in a given ecosystem, $i = 1, 2, \dots, n$.

Together with (1), we consider its δ-perturbations of the form

$$\dot{x}_i = x_i g_i(x), \ \ x \in X, \quad \text{such that} \quad |g_i(x) - f_i(x)| < \delta \ \text{ for all } x \in X. \tag{2}$$

It is of course assumed that $g : X \to \mathbb{R}^n$ is a continuous function and $\sum_i x_i g_i(x) = 0$ for each $x \in X$. We assume further that both (1) and (2) have the uniqueness

property. Denote by $\Phi(\cdot, x)$ and $\Psi(\cdot, x)$ the solutions of (1) and (2) starting in $x \in X$. It is immediate that both $\Phi : \mathbb{R} \times X \to X$ and $\Psi : \mathbb{R} \times X \to X$ are dynamical systems on X.

The boundary of X is denoted by Y. Y is invariant under Φ and Ψ. System (1) is called *permanent* (or uniformly persistent) if Y is a repeller. In ecological equations, permanence means the ultimate survival of all species. If (2) is permanent, then $(\mathcal{A}_\Psi, Y)$ forms an attractor–repeller pair, where $\mathcal{A}_\Psi$ denotes the maximal compact Ψ-invariant set in $X \setminus Y$.

The aim of this section is to give a sufficient condition for *robust permanence*, guaranteeing that every system near (1) is permanent.

DEFINITION 2.1. *Let us call a continuous mapping $P : \mathbb{R}^n_+ \to \mathbb{R}$ a* good average Liapunov function *(GALF)* for (1) *if*

(a) *$P(x) = 0$ for all $x \in \partial\mathbb{R}^n_+$, $P(x) > 0$ for all $x \in \operatorname{int}\mathbb{R}^n_+$;*

(b) *P is differentiable on $\operatorname{int}\mathbb{R}^n_+$ and $p_i(x) := \frac{x_i}{P(x)}\frac{\partial P}{\partial x_i}$ can be extended to a continuous function on X for every i;*

(c) *for every $y \in Y$ there is a positive constant T_y with the property that*

$$\int_0^{T_y} \sum_i p_i(\Phi(t,y)) f_i(\Phi(t,y))\, dt > 0.$$

Now we are in a position to present the main result of this section.

THEOREM 2.2. *If there is a GALF for* (1)*, then* (1) *is robustly permanent: There are a $\delta > 0$ and a compact subset S of $X \setminus Y$ such that every δ-perturbation* (2) *of* (1) *is permanent and $\mathcal{A}_\Psi$ is contained in S.*

Remark 2.3. If the inequality in (c) is reversed, then Y can be shown to be a robust attractor for (1).

The concept of an *average Liapunov function* (ALF) *for* (1) (with (a), (c), and a weaker version of (b), namely, the assumption that

$$\text{(3)} \qquad \text{the function} \quad \frac{\dot P}{P} = \sum_{i=1}^n \frac{1}{P(x)}\frac{\partial P}{\partial x_i} x_i f_i(x) \quad \text{is continuous on } X \ )$$

and Theorem 2.2 (without the robustness conclusion) are due to Hofbauer [27], inspired by Schuster, Sigmund, and Wolff [58]. The standard candidate for an ALF satisfying (a) and (b) is $P(x) = \prod_{i=1}^n x_i^{p_i}$ with constants $p_i > 0$.[1] In this case $p_i(x) = p_i$, and Theorem 2.2 reduces to the following.

COROLLARY 2.4. *Suppose there are positive constants p_i, $i = 1, \dots, n$, such that for each $y \in Y$ of* (1) *there is a time $T_y > 0$ such that $\int_0^{T_y} \sum_i p_i f_i(\Phi(t,y))\, dt > 0$. Then* (1) *is robustly permanent.*

The concept of an ALF is—like that of a Liapunov function—a topological one: It can be formulated [37] in metric spaces X to show that a closed invariant subset Y is a repeller. The concept of a GALF, on the other hand, makes use of the smooth structure of X. Besides for the simplex, it applies to X being any manifold with corners (i.e., modeled after $\mathbb{R}^n_+$). Theorem 2.2 continues to hold in this more general

[1] In most practical applications this function has been used. Hutson [37] and Hofbauer [28] use more general ALFs (that are not GALFs, however). But in these instances, the standard form $P(x) = \prod_{i=1}^n x_i^{p_i}$ would be sufficient if used as in Theorem 5.5 below, i.e., taking different choices of the vector p on different Morse sets.

setting as long as X is compact. A simple example, arising in section 11.5, is X being a product of simplices. For another simple example let X be a manifold with smooth boundary, as in section 11.1. In this case, the standard GALF is simply the distance to the boundary manifold Y. This standard GALF is even good enough here to *characterize* robust repulsivity of Y.

If the state space is not compact, then some adjustments have to be made. We describe the most important case of (1) defining an ecological differential equation on $\mathbb{R}^n_+$. We restrict ourselves to systems (1) that generate dissipative (semi)flows. Φ need no longer define a flow on $\mathbb{R}^n_+$ (solutions need not be defined for all negative times, as, e.g., in the logistic equation $\dot{x} = x(1-x)$ on $\mathbb{R}_+$). Let X be a compact absorbing subset for the local flow Φ. Then $Y = X \cap \partial\mathbb{R}^n_+$ is also compact and forward invariant under (1). There are at least two ways to define robustness:

1. Consider δ-perturbations only on X and assume that X remains absorbing for the perturbed flow Ψ, i.e., $\Psi(t, X) \subset X$ holds for all $t \geq 0$. This is done in [26, Cor. 4.6], where X is taken as a cube.
2. Allow perturbations of f in (1) in the strong Whitney topology, an approach taken in [57].

Either way, Theorem 2.2 and Corollary 2.4 remain true as stated.

Corollary 2.4 (again without the robustness conclusion) has been widely used to prove permanence of population dynamical systems; see Hofbauer and Sigmund [35]. The new aspect treated in this paper and the difference between ALF and GALF is illustrated by the following example. For a different kind of robustness, see [38].

Example 2.5. Consider $n = 2$, so that (1) is of the form $\dot{x}_1 = x_1(1-x_1)F(x_1)$, $\dot{x}_2 = -x_2(1-x_2)F(1-x_2)$. Both systems

$$\text{(i)} \begin{cases} \dot{x}_1 = x_1(1-x_1)(1/2-x_1), \\ \dot{x}_2 = -x_2(1-x_2)(x_2-1/2) \end{cases} \quad \text{and} \quad \text{(ii)} \begin{cases} \dot{x}_1 = x_1^2(1-x_1)^2(1/2-x_1), \\ \dot{x}_2 = -x_2^2(1-x_2)^2(x_2-1/2) \end{cases}$$

are permanent, since $Y = \{(0,1)\} \cup \{(1,0)\}$ is a repeller. However, (i) is robustly permanent, whereas (ii) is not. The reason is of course that in (i) both $\{(0,1)\}$ and $\{(1,0)\}$ are hyperbolic, but they are not for (ii). This is captured by the auxiliary function $P(x_1, x_2) = x_1 x_2$ (or $x_1^{p_1} x_2^{p_2}$ for any $p_1, p_2 > 0$): Condition (c) reduces to $\min\{F(0), -F(1)\} > 0$, which holds for (i) but not for (ii). Hence P is a GALF for (i) but not (ii). On the other hand, the new auxiliary function $\tilde{P}(x_1, x_2) = e^{-1/x_1 - 1/x_2}$ satisfies

$$x_1(1-x_1)^2\left(\frac{1}{2}-x_1\right)\tilde{p}_1(x_1,x_2) - x_2(1-x_2)^2\left(x_2-\frac{1}{2}\right)\tilde{p}_2(x_1,x_2) = \frac{(x_1-x_2)^2}{2}$$

for each $(x_1, x_2) \in X \setminus Y$. Taking continuous extensions, we see that (a), (c), and (3) are satisfied for (ii). But (b) is violated since $\tilde{p}_1(x_1, x_2) = 1/x_1$ and $\tilde{p}_2(x_1, x_2) = 1/x_2$ do not have continuous extensions to X. Thus $\tilde{P}$ is not a GALF (but only an ALF) for (ii).

Next we illustrate the method of GALFs by deriving a stability criterion for a heteroclinic cycle. For further examples, see [27] for the planar case and [31] for higher-dimensional examples.

Example 2.6. Consider the replicator dynamics

$$\dot{x}_i = x_i((Ax)_i - xAx), \quad i = 1, 2, \ldots, n, \tag{4}$$

on the simplex $X = \{x \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^n x_i = 1\}$, with $n = 3$ for a rock-scissors-paper game with payoff matrix

$$A = \begin{pmatrix} 0 & -a_2 & b_3 \\ b_1 & 0 & -a_3 \\ -a_1 & b_2 & 0 \end{pmatrix}. \tag{5}$$

Then the boundary Y forms a heteroclinic cycle, with "outgoing" eigenvalues $b_i > 0$ and "incoming" eigenvalues $-a_i < 0$ at the ith corner. Consider the standard function $P(x) = \prod_{i=1}^3 x_i^{p_i}$ (with $p_i > 0$ to be suitably chosen), which satisfies (a) and (b). Since every orbit on the boundary Y converges to one of the corners, it is sufficient to check (c) at these three equilibria. Hence (c) leads to a system of three linear inequalities

$$b_1 p_2 > a_1 p_3, \quad b_2 p_3 > a_2 p_1, \quad b_3 p_1 > a_3 p_2, \tag{6}$$

which can be summarized as

$$A^T p > 0 \quad \text{for suitable} \quad p > 0. \tag{7}$$

Obviously (6) has a solution in $p_i > 0$ if and only if

$$b_1 b_2 b_3 > a_1 a_2 a_3. \tag{8}$$

In this case, by Corollary 2.4, (4) is robustly permanent, i.e., the heteroclinic cycle Y is robustly repelling. If the inequalities in (6), (7), or, equivalently, in (8) are reversed, then, by Remark 2.3, the heteroclinic cycle Y is robustly attracting for (4). Note that the result does not depend on the special dynamics (4) but only on the "external eigenvalues" at the three corner equilibria which correspond to the entries of the matrix A. (Note that the above derivation of the stability criterion (8) using GALFs is much easier compared to other methods, such as finding a true Liapunov function near Y or applying Poincaré sections [35].)

Now we turn to the proof of Theorem 2.2. We shall make use of Corollary 6.1.2 of [8], which is a reformulation of Theorem 9 of the 1957 Russian edition[2] of Zubov's monograph [68].

ZUBOV–URA–KIMURA THEOREM. *Let (W, d) be a locally compact separable metric space and let Θ be a dynamical system on W. Finally, let $\emptyset \neq M$ be a compact isolated Θ-invariant set in W. Suppose that M is not a repeller. Then $\emptyset \neq \omega(x) \subset M$ for some $x \notin M$.*

Neither Zubov's work [68] nor the paper by Ura and Kimura [64] had been generally known before the 1970 monograph of Bhatia and Szegö [8]. Had they been known before, they might have led to essential simplifications in establishing such important notions of topological dynamics as the Auslander–Seibert duality between stability

[2]The proof of Theorem 9 in [68] is based on Theorem 7 of that work. Unfortunately, this latter statement is false. As it is remarked in the 1964 English edition [69], the error was pointed out by S. Lefschetz to V. I. Zubov. A corrected version of Theorem 7 was published by Bass [5], an associate of Lefschetz. A corrected version of Theorem 7 appears also in [69] and (although the last sentence on p. 35 of [69] is still false) makes the derivation of Theorem 9 correct, too. From Theorem 8 onward, section 11 of the English edition is a word-for-word translation of the Russian edition and contains several interconnected results on the local behavior of continuous-time dynamical systems near compact isolated invariant sets. More or less the same set of results was obtained by Ura and Kimura [64] independently in 1960. What we call the Zubov–Ura–Kimura theorem is a collection of several technical lemmas of [64, pp. 26–31].

and boundedness, Bhatia's concept of weak attraction, and the Wilson–Yorke hyperbolic Liapunov function in the early sixties.[3] A large number of much later results in persistence theory during the eighties (including the Butler–McGehee lemma in the appendix of [15]) followed easily from those in [68] and [64].

Proof of Theorem 2.2. Suppose P is a GALF for (1). We claim that there are three constants $c, \delta, T > 0$ and a compact subset S of $X \setminus Y$ with the following property. Given $x \in \mathrm{cl}(X \setminus S) \setminus Y$ arbitrarily, there exists a time $T_x \in (0, T]$ such that

$$P(\Psi(T_x, x)) > (1 + c)P(x) \quad \text{for every } \delta\text{-perturbation (2) of (1).} \tag{9}$$

In fact, condition (c) plus an easy compactness argument imply that there are positive constants c, d, and T such that for all $x \in I(d) = \{x \in X : P(x) \le d\}$ (= a small compact neighborhood of Y) there is a time $T_x \in (0, T]$ with

$$\int_0^{T_x} \sum_i p_i(\Phi(t, x)) f_i(\Phi(t, x))\, dt > 3c > 0. \tag{10}$$

Uniform continuity of $p_i f_i$, $i = 1, 2, \ldots, n$, provides an $\varepsilon > 0$ such that

$$|p_i(z)f_i(z) - p_i(w)f_i(w)| < \frac{c}{nT} \quad \text{whenever} \quad z, w \in X \text{ and } |z - w| < \varepsilon.$$

Since the p_i's are bounded and $|g - f| < \delta$, we obtain for δ small enough by the triangle inequality that

$$|p_i(z)g_i(z) - p_i(w)f_i(w)| < \frac{2c}{nT} \quad \text{whenever} \quad z, w \in X \text{ and } |z - w| < \varepsilon.$$

By a standard Arzelà–Ascoli argument, the uniqueness property of (1) implies there is a $\delta > 0$ such that $|\Psi(t, x) - \Phi(t, x)| < \varepsilon$ for $t \in [0, T]$, $x \in X$, and every δ-perturbation (2) of (1). In view of inequality (10), we conclude via condition (b) that

$$\log(P(\Psi(T_x, x)) - \log(P(x)) = \int_0^{T_x} \sum_i p_i(\Psi(t, x)) g_i(\Psi(t, x))\, dt$$

$$\ge - \int_0^{T_x} \left| \sum_i p_i(\Psi(t, x)) g_i(\Psi(t, x)) - \sum_i p_i(\Phi(t, x)) f_i(\Phi(t, x)) \right| dt$$

$$+ \int_0^{T_x} \sum_i p_i(\Phi(t, x)) f_i(\Phi(t, x))\, dt \ge -\frac{2T_x c}{T} + 3c \ge c \qquad \text{for each } x \in I(d) \setminus Y.$$

Set $S = \mathrm{cl}(X \setminus I(d))$ and note that $\mathrm{cl}(X \setminus S) = I(d)$. Since $e^c > 1 + c$, inequality (9) follows.

[3] All proofs of the Zubov–Ura–Kimura theorem work equally well for discrete-time dynamical systems. However, the first discrete-time version of the Zubov–Ura–Kimura theorem was discovered independently of [68] and [64]. It is Lemma 1 in Browder [10] (termed "a crucial one" by Browder himself), stating that a strongly ejective fixed point is repulsive. Establishing his famous existence theorem on nonejective fixed points, Browder worked out a great deal of basic topological dynamics using his own terminology. His crucial lemma is a direct consequence of the (discrete-time semidynamical version of the) Zubov–Ura–Kimura theorem. (Multivalued and various discretization aspects are investigated in [61] and [22], respectively.)

Next we point out that

$$I(d) \setminus \gamma_\Psi^+(x) \neq \emptyset \qquad \text{for each} \quad x \notin Y. \tag{11}$$

In fact, suppose there is a $z \in I(d)$ such that $\gamma_\Psi^+(z) \subset I(d)$. Then P attains its maximum value on $\mathrm{cl}(\gamma_\Psi^+(z))$ at some point w. In particular, if $z \notin Y$, this implies $P(w) > 0$ and the existence of a time sequence $\{\tau_n\} \subset \mathbb{R}_+$ such that $P(\Psi(\tau_n, z)) \to P(w)$ as $n \to \infty$. Applying (9), we obtain that $P(\Psi(T_{\Psi(\tau_n,z)}, \Psi(\tau_n, z))) \geq (1 + c)P(\Psi(\tau_n, z)) \to (1 + c)P(w)$, a contradiction to the choice of w.

As a byproduct of (11), $I(d)$ is an isolating neighborhood of Y and for each $x \in I(d) \setminus Y$, inclusion $\emptyset \neq \omega_\Psi(x) \subset Y$ is impossible. By the Zubov–Ura–Kimura theorem, Y is a repeller for (2) and the dual attractor $\mathcal{A}_\Psi$ is contained in S. □

Remark 2.7. As for any proof using compactness considerations, the proof of Theorem 2.2 is also nonconstructive. However, it is not hard to see that all "intrinsically nonconstructive ingredients" of the proof are contained in assumption (c). To be more precise, assume that the conditions of Theorem 2.2 are all satisfied. In addition, assume that

(H1) there exist positive constants c, T with the property that, given $y \in Y$ arbitrarily, $\int_0^{T_y} \sum_i p_i(\Phi(t, y)) f_i(\Phi(t, y))\, dt > 4c$ for some $T_y \in (0, T]$.

Finally, assume that (no extra assumptions on the g_i's are needed!)

(H2) the functions p_i, f_i, $i = 1, 2, \ldots, n$, are (globally) Lipschitz.

Reconsidering the proof of Theorem 2.2, it is routine to check that all compactness arguments including the Zubov–Ura–Kimura argument can be replaced by Gronwall inequalities. The final conclusion is that the parameters δ and the distance of S from Y are both larger than $\Lambda c \exp(-\lambda T)$, where $\lambda, \Lambda > 0$ are computable constants and do not depend on c, T (provided by (H1)) and on the perturbation g, but only on the various Lipschitz constants (provided by (H2)). Hence the GALF assumption, together with (H1) and (H2), provides a way of estimating the distance between S and Y. Thus we have a feasible approach to the problem of "practical persistence" discussed by Hutson and Mischaikow [39] in two dimensions.

3. Exponential repulsivity. In this section we explore the concept of an average Liapunov function and its relation to exponential repulsion and existence of (ordinary) Liapunov functions.

THEOREM 3.1. (1) *If P is an ALF for* (1)*, then there exist an open neighborhood $\mathcal{N}$ of Y in X and positive constants κ_1, κ_2 such that*

$$P(\Phi(t, x)) \leq \kappa_1 e^{\kappa_2 t} P(x) \quad \text{for each} \quad x \in \mathcal{N} \quad \text{and} \quad t \leq 0. \tag{12}$$

(2) *If P is a GALF for* (1)*, then there exist an open neighborhood $\mathcal{N}$ of Y in X and positive constants $\delta, \kappa_1, \kappa_2$ such that for each δ-perturbation*

$$P(\Psi(t, x)) \leq \kappa_1 e^{\kappa_2 t} P(x) \quad \text{for each} \quad x \in \mathcal{N} \quad \text{and} \quad t \leq 0.$$

(3) *If $P(x) = \prod_{i=1}^n x_i^{p_i}$ is a GALF for* (1) *and letting $\delta > 0$ be the same constant as in Theorem* 2.2*, then there exist an open neighborhood $\mathcal{N}$ of Y in X and positive constants $\kappa_1, \kappa_2, \kappa_3$ such that*

$$d_E(\Psi(t, x), Y) \leq \kappa_1 e^{\kappa_2 t} (d_E(x, Y))^{\kappa_3} \quad \text{for each} \quad x \in \mathcal{N} \quad \text{and} \quad t \leq 0.$$

Here $d_E(x, Y)$ denotes the Euclidean distance between a point $x \in X$ and the set Y.

Proof. The proof of assertion 1 will be omitted since it is the same as that of 2 but ignores the robustness.

(2) The application of the Zubov–Ura–Kimura theorem in the last step of proving Theorem 2.2 will be replaced by an explicit computation, as in the earlier proofs in [27], [37]. Parameters introduced and auxiliary inequalities derived in proving Theorem 2.2 will be used throughout.

Since $\sum |p_i g_i| \leq \kappa$ for some $\kappa > 0$, we obtain via integrating the identity

$$\frac{d}{dt} \log P(\Psi(t,x)) = \sum_{i=1}^{n} p_i(\Psi(t,x)) g_i(\Psi(t,x)) \quad \text{for } t \in \mathbb{R} \text{ and } x \in X \setminus Y, \tag{13}$$

a consequence of assumption (b) that

$$e^{\kappa\tau} P(x) \geq P(\Psi(\tau,x)) \geq e^{-\kappa\tau} P(x) \quad \text{for every} \quad x \in X\,,\ \tau \geq 0 \tag{14}$$

and every δ-perturbation (2) of (1).

Suppose now that $x \in I(d) \setminus Y$. Since $\mathcal{A}_\Psi \subset S = \mathrm{cl}(X \setminus I(d))$, there exists a nonnegative integer $K(x)$ with the properties that $\Psi([0, K(x)T], x) \subset I(d)$ but $\Psi(t,x) \notin I(d)$ for some $t \in (K(x)T, (K(x)+1)T]$. Set $T_{x,0} = 0$ and, recursively, as long as $T_{x,k} \leq K(x)T$, set $T_{x,k+1} = T_{\Psi(T_{x,k},x)}$, $k = 0, 1, \ldots,$ (say) $k(x)$. Inequality (9) can be iterated $k(x)$ times and yields that

$$P(\Psi(T_{x,k}, x)) \geq (1+c)^k P(x) \quad \text{for each} \quad k = 0, 1, \ldots, k(x).$$

Recall that $0 < T_{x,k+1} - T_{x,k} \leq T$. In view of inequality (14), it follows immediately that

$$P(\Psi(t,x)) \geq e^{-\kappa T}(1+c)^k P(x) \quad \text{whenever } \ T_{x,k} \leq t \leq T_{x,k+1}\,,$$

and $k = 0, 1, \ldots, k(x)$. By using $T_{x,k+1} \leq (k+1)T$, we conclude that

$$P(\Psi(t,x)) \geq \frac{e^{-\kappa T}}{1+c} \cdot (1+c)^{t/T} \cdot P(x) \quad \text{whenever} \quad t \in [0, K(x)T]. \tag{15}$$

Choose $T^* > 0$ in such a way that $e^{-\kappa T} \cdot (1+c)^{-1+T^*/T} > 1$ and set $\Delta = e^{-\kappa T^*} d$. We claim that

$$P(\Psi(t, I(\Delta))) \leq d \qquad \text{for each} \quad t \leq 0. \tag{16}$$

Suppose this is not the case. Then there exist a $t^* > 0$ and an $x^* \in X$ with $P(x^*) \leq \Delta$, $P(\Psi(-t^*, x^*)) = d$ but $P(\Psi(t, x^*)) < d$ for each $t \in (-t^*, 0]$. By the construction,

$$e^{-\kappa t^*} d = e^{-\kappa t^*} P(\Psi(-t^*, x^*)) \leq P(\Psi(t^*, \Psi(-t^*, x^*))) = P(x^*) \leq \Delta,$$

and thus $t^* \geq T^*$. A similar application of (14) and the simple inequality $T < T^*$ show that

$$P(\Psi([0,T], x^*)) \leq e^{\kappa T} P(x^*) < e^{\kappa T^*} \Delta = d.$$

We conclude that $K(\Psi(-t^*, x^*))T \geq t^* \geq T^*$, and hence, by using inequality (15) with $t = T^*$ and $x = \Psi(-t^*, x^*)$,

$$P(\Psi(-t^* + T^*, x^*)) \geq e^{-\kappa T} \cdot (1+c)^{-1+T^*/T} \cdot P(\Psi(-t^*, x^*)) > 1 \cdot d = d,$$

a contradiction.

Set $\mathcal{N} = \{x \in X \mid P(x) < \Delta\}$. By virtue of (16), we can pass to negative times and obtain from inequality (15) that

$$P(\Psi(t,x)) \le \frac{1+c}{e^{-\kappa T}} \cdot (1+c)^{t/T} \cdot P(x) \quad \text{whenever} \quad x \in \mathcal{N} \ \text{ and } \ t \le 0. \tag{17}$$

This completes the proof of assertion 2.

(3) The Euclidean distance between a point $x \in X$ and the set Y equals

$$d_E(x,Y) = \left(\min \left\{ \sum_{j=1}^{n} (x_j - y_j)^2 \mid y \in Y \right\} \right)^{1/2} = \min_{1 \le j \le n} x_j.$$

By compactness, there exist continuous, strictly increasing functions $\alpha, \beta : [0, 1/n] \to \mathbb{R}^+$ with the properties that $\alpha(0) = \beta(0) = 0$ and

$$\alpha(d_E(x,Y)) \le P(x) \le \beta(d_E(x,Y)) \quad \text{whenever} \quad x \in X. \tag{18}$$

Combining (16) and (18), the rate of repulsion near Y can be estimated *in terms of the Euclidean distance function.*

For example, the standard GALF $P(x) = \prod_{i=1}^{n} x_i^{p_i}$ satisfies

$$(d_E(x,Y))^{\sum_{i=1}^n p_i} = \prod_{i=1}^{n} \left(\min_{1 \le j \le n} x_j \right)^{p_i} \le P(x) \le \min_{1 \le i \le n} x_i^{p_i} \le (d_E(x,Y))^{\min_{1 \le i \le n} p_i}$$

for each $x \in X$ and leads to the desired exponential rate of repulsion. In fact, given $x \in \mathcal{N}$ arbitrarily, we obtain that

$$(d_E(\Psi(t,x),Y))^{\sum_{i=1}^n p_i} \le \frac{1+c}{e^{-\kappa T}} \cdot (1+c)^{t/T} \cdot (d_E(x,Y))^{\min_{1 \le i \le n} p_i}$$

for each $t \le 0$. This shows how constants $\kappa_1, \kappa_2, \kappa_3$ in assertion 3 must be chosen. □

COROLLARY 3.2. *If P is an ALF for* (1)*, then there exists an exponentially increasing Liapunov function for* (1)*. In other words, there exist a negatively invariant open neighborhood $\mathcal{U}$ of Y in X, a positive constant κ, and a continuous function $V : \mathcal{U} \to \mathbb{R}_+$ such that $V(x) = 0$ if and only if $x \in Y$ and*

$$V(\Phi(t,x)) \le e^{\kappa t} V(x) \quad \textit{for each} \quad x \in \mathcal{U} \ \textit{ and } \ t \le 0. \tag{19}$$

Proof. The standard integration trick [8] is used for eliminating κ_1 from (12). We fix a negatively invariant open neighborhood $\mathcal{U}$ of Y in $\mathcal{N}$ and define

$$V(x) = \int_{-\infty}^{0} e^{-\kappa_2 t/(1+\Delta)} P(\Phi(t,x)) dt \quad \text{for each} \quad x \in \mathcal{U}. \tag{20}$$

Here $\Delta > 0$ is arbitrary and $\kappa = \kappa_2/(1+\Delta)$ in (19). □

LEMMA 3.3. *If P is a GALF for* (1) *and $W = \exp(w)$ is any positive C^1 function, then $\tilde{P} = PW$ is also a GALF for* (1)*.*

Proof. $\tilde{P}$ obviously satisfies condition (a) in Definition 2.1. (b) follows from $\tilde{p}_i := \frac{x_i}{\tilde{P}(x)} \frac{\partial \tilde{P}}{\partial x_i} = p_i + x_i \frac{\partial w}{\partial x_i}$, $i = 1, 2, \ldots, n$. And the identity

$$\frac{1}{T} \int_0^T \sum_i \left(\tilde{p}_i(\Phi(t,y)) - p_i(\Phi(t,y)) \right) f_i(\Phi(t,y)) \, dt = \frac{w(\Phi(T,y)) - w(y)}{T}, \quad y \in Y,$$

together with Lemma 4.2 below, shows (c). □

THEOREM 3.4. *For $P(x) = \prod_{i=1}^n x_i^{p_i}$ (with $p_i > 0$) and $f \in C^1(X, \mathbb{R}^n)$ the following conditions are equivalent:*

(A) *$P(x)$ is a GALF for (1).*

(B) *There exist a negatively invariant open neighborhood $\mathcal{U}$ of Y in X, a positive constant κ, and a C^1 function $W : \mathcal{U} \to (0, \infty)$ such that $V = PW$ is an exponentially increasing Liapunov function for (1):*

$$\dot{V}(x) \geq \kappa V(x) \quad \text{for all} \quad x \in \mathcal{U}. \tag{21}$$

Proof. If (B) holds, then V is a GALF, and hence by Lemma 3.3, with $P, W, \tilde{P}$ replaced by $V, 1/W, P$, also P must be a GALF. Now suppose that $P(x) = \prod_{i=1}^n x_i^{p_i}$ is a GALF for (1). We write $\Phi_i(t,x) = x_i Q_i(t,x)$. By Corollary 6.1, $Q_i(t,x) > 0$ for all $t \in \mathbb{R}$ and $x \in X$. Define $q(t,x) := \sum_i p_i \log Q_i(t,x)$. Then (12) implies

$$P(\Phi(t,x)) = e^{q(t,x)} P(x) \leq \kappa_1 e^{\kappa_2 t} P(x) \quad \text{for } t \leq 0. \tag{22}$$

Furthermore $\frac{\partial q(t,x)}{\partial t} = \sum_i p_i f_i(\Phi(t,x)) =: \tilde{f}(\Phi(t,x))$, and for the partial derivatives

$$\frac{\partial}{\partial t}\frac{\partial q}{\partial x_j}(t,x) = \sum_i \frac{\partial \tilde{f}}{\partial x_i}(\Phi(t,x)) \frac{\partial \Phi_i}{\partial x_j}(t,x). \tag{23}$$

Let L be a Lipschitz constant of (1). Then Gronwall's inequality implies $|\frac{\partial \Phi_i}{\partial x_j}(t,x)| \leq e^{L|t|}$, and hence in (23)

$$\left| \frac{\partial}{\partial t}\frac{\partial q}{\partial x_j}(t,x) \right| \leq C e^{L|t|}. \tag{24}$$

After integration this gives

$$\left| \frac{\partial q}{\partial x_j}(t,x) \right| \leq C' e^{L|t|} \tag{25}$$

for some positive constants C, C'. Now use P^α (for any $\alpha > 0$) instead of P in (20) and consider

$$V_\alpha(x) = \int_{-\infty}^0 e^{-\alpha\kappa_2 t/(1+\Delta)} P(\Phi(t,x))^\alpha dt = P(x)^\alpha W_\alpha(x) \tag{26}$$

with

$$W_\alpha(x) = \int_{-\infty}^0 e^{-\alpha\kappa_2 t/(1+\Delta)} e^{\alpha q(t,x)} dt. \tag{27}$$

Then, by (22), for every $\alpha > 0$ and $\Delta > 0$, the function V_α is continuous on X and satisfies (19) with $\kappa = \alpha\kappa_2/(1+\Delta)$. The function W_α is continuous and positive on X. Formal differentiation of (27) gives

$$\frac{\partial W_\alpha}{\partial x_j}(x) = \int_{-\infty}^0 e^{-\alpha\kappa_2 t/(1+\Delta)} e^{\alpha q(t,x)} \alpha \frac{\partial q}{\partial x_j}(t,x) dt. \tag{28}$$

With (22) and (25) we can estimate the integrand up to a constant factor by $\exp(-\frac{\alpha\kappa_2 t}{1+\Delta} + \alpha\kappa_2 t - Lt)$ (for $t < 0$). Hence for

$$\alpha > \frac{1+\Delta}{\Delta}\frac{L}{\kappa_2} \tag{29}$$

the indefinite integral in (28) converges absolutely and uniformly in $x \in X$. This implies that for all these α large enough, W_α is C^1. Then the claim follows for $V = V_\alpha^{1/\alpha} = PW_\alpha^{1/\alpha} =: PW$. □

Remark 3.5. We note that W (and hence V) can be made as smooth as the vector field (1) by choosing α sufficiently large: This follows easily by further differentiating (28). The relationship between (12) and (21) is in line with the general observation that, under certain conditions, inequalities can be differentiated with respect to parameters [9], [47].

Remark 3.6. Lemma 3.3 shows that to each GALF P there belongs a whole equivalence class of GALFs differing by a smooth positive factor. Theorem 3.4 shows that for a standard GALF, there is a true Liapunov function among these equivalent GALFs. Still, the advantage of the GALF concept is its considerably easier practical applicability, as compared to a true Liapunov function. Finding a standard GALF for (1) is reduced in the next section to the algebraic problem of finding suitable constants $p_i > 0$. In the setting of manifolds with smooth boundary in section 11.1, there is essentially a unique standard GALF, which is simply the distance to the boundary manifold. The GALF conditions for this simple function *characterize* robust repulsivity of the boundary. Finding an explicit true Liapunov function is considerably more difficult.

Remark 3.7. Theorem 3.4 (B) implies another, very simple proof of robust permanence (under the stronger assumption $f \in C^1$): Write $V = Pe^w$. Then along interior solutions of a δ-perturbation (2)

$$\dot{V}/V = \sum_i p_i \frac{\dot{x}_i}{x_i} + \dot{w} = \sum_i \left(p_i + x_i \frac{\partial w}{\partial x_i}\right) g_i(x)$$

$$= \sum_i \left(p_i + x_i \frac{\partial w}{\partial x_i}\right) f_i(x) + \sum_i \left(p_i + x_i \frac{\partial w}{\partial x_i}\right) (g_i(x) - f_i(x)).$$

The first sum is $\geq \kappa > 0$ by (21) and the second term is less than a constant (since w is C^1) times δ. Hence for δ small enough, V is a local Liapunov function near Y also for (2).

4. GALF and minimax. We need the minimax theorem in the following simplified formulation (see, e.g., [59]).

MINIMAX THEOREM. *Let A, B be Hausdorff topological vector spaces and let $\Gamma : A \times B \to \mathbb{R}$ be a continuous bilinear function. Finally, let C and D be nonempty, convex, compact subsets of A and B, respectively. Then*

$$\min_{a\in C}\max_{b\in D}\Gamma(a,b) = \max_{b\in D}\min_{a\in C}\Gamma(a,b).$$

In what follows $\mathcal{M}_\Phi$ denotes the collection of Φ-invariant Borel probability measures on Y. The collection of all Borel probability measures on Y is denoted by $\mathcal{M}$. Recall that both $\mathcal{M}_\Phi$ and $\mathcal{M}$ are nonempty, convex, weakly-$*$ compact subsets of

$C_w^*(Y,R)$, the dual space of $C(Y,R)$ equipped with the weak-$*$ topology. The subset $\mathcal{M}_\Phi^E$ of ergodic Φ-invariant measures is the set of extreme points of $\mathcal{M}_\Phi$. The characteristic function of a Borel set $B \subset Y$ is denoted by χ_B.

LEMMA 4.1.[4] *Let $h : Y \to \mathbb{R}$ be a continuous function. Then*

$$\min_{\mu\in\mathcal{M}_\Phi} \int_Y h\, d\mu = \min_{y\in Y} \left\{ \limsup_{T\to\infty} \frac{1}{T} \int_0^T h(\Phi(t,y))\, dt \right\}. \tag{30}$$

Proof. Replacing "lim" by "lim sup" in the ergodic theorem, we have that

$$\int_Y h\, d\mu = \limsup_{T\to\infty} \frac{1}{T} \int_0^T h(\Phi(t,y))\, dt \quad \text{for } \mu\text{–almost all } y \in Y. \tag{31}$$

Note that the right-hand side of (31) defines a lower semicontinuous function on Y. Since Y is compact and $\mathcal{M}_\Phi$ is weakly-$*$ compact, we can take minima on both sides. This proves the "$\geq$ part" of (30).

To prove the "$\leq$ part," we argue via reductio ad absurdum and suppose there exist a $y_0 \in Y$, an $\varepsilon_0 > 0$, and a time sequence $\{\tau_n\} \subset \mathbb{R}_+$ such that $\tau_n \to \infty$ and

$$\int_Y h\, d\mu > \varepsilon_0 + \frac{1}{\tau_n} \int_0^{\tau_n} h(\Phi(t,y_0))\, dt \quad \text{for each } \mu \in \mathcal{M}_\Phi \text{ and } n = 1,2,\ldots. \tag{32}$$

By letting $\mu_n(B) = \frac{1}{\tau_n}\int_0^{\tau_n} \chi_B(\Phi(t,y_0))\, dt$ for each Borel set $B \subset Y$, a $\mu_n \in \mathcal{M}$ is defined and the inequality in (32) goes over into $\int_Y h\, d\mu > \varepsilon_0 + \int_Y h\, d\mu_n$. We may assume that, in the weak-$*$ topology, $\mu_n \to \mu_0$ for some $\mu \in \mathcal{M}$. The crucial observation is that $|\mu_n(\Phi(\tau,B)) - \mu_n(B)| \leq \frac{2|\tau|}{\tau_n}$ for each Borel set $B \subset Y$, $\tau \in \mathbb{R}$ and $n = 1,2,\ldots$. By letting $n \to \infty$, we conclude that[5] $\mu_0 \in \mathcal{M}_\Phi$. Hence $\int_Y h\, d\mu_0 \geq \varepsilon_0 + \int_Y h\, d\mu_0$, a contradiction. ∎

LEMMA 4.2. *For any continuous function $h : Y \to \mathbb{R}$, the following properties are equivalent:*

(i) $\min_{y\in Y}\{\limsup_{T\to\infty} \frac{1}{T}\int_0^T h(\Phi(t,y))\, dt\} > 0$.

(ii) *For every $y \in Y$ there is a $T(y) > 0$ with $\int_0^{T(y)} h(\Phi(t,y))\, dt > 0$.*

Proof. (i) $\Rightarrow$ (ii) is trivial. Suppose now that (ii) is satisfied. By an easy compactness argument, we see there is no loss of generality in assuming there are positive constants c_0, T_1, T_2 such that

$$T_1 \leq T(y) \leq T_2 \quad \text{and} \quad \int_0^{T(y)} h(\Phi(t,y))\, dt > c_0 \quad \text{for all } y \in Y.$$

Set $\tau_0 = 0$ and, recursively, $\tau_n(y) = \tau_{n-1}(y) + T(\Phi(\tau_{n-1}(y), y))$. By the construction, $nT_1 \leq \tau_n \leq nT_2$ and

$$\frac{1}{\tau_n(y)} \int_0^{\tau_n(y)} h(\Phi(t,y))\, dt \geq \frac{c_0}{T_2} \qquad \text{for each } n = 1,2,\ldots.$$

Even with lim sup replaced by lim inf, (i) follows immediately. ∎

[4] This is Exercise 8.5 on p. 57 in [51] (with "limsup" replaced by "liminf"). Proofs were written in [30] and [56], who derived it from a much more general setting. Other generalizations were given in [3] and [11]. We include a proof for completeness.

[5] The argumentation leading to μ_0 is truly fundamental and plays a vital role in the ergodic theory of dynamical systems from its very beginnings in Krylov and Bogoliubov [46] to (proving the first part of) Theorem 4.3 of Schreiber [57].

Remark 4.3. Note that condition (c) in Definition 2.1 has to be checked only for $y \in MC_\Phi(Y)$, the minimal center of attraction of Y (i.e., by definition, the smallest compact Φ-invariant set containing the support of each invariant measure in $\mathcal{M}_\Phi(Y)$). This follows from a twofold application of Lemmas 4.1 and 4.2 with Y and Y replaced by $MC_\Phi(Y)$; cf. [56]. Note that $MC_\Phi(Y)$ is contained in the closure of the set of all $\Phi|_Y$-recurrent points.

THEOREM 4.4. *The following properties are equivalent:*

(α) *For some* $p_1, p_2, \dots, p_n > 0$ *suitably chosen,* $P(x) = \prod_{i=1}^n x_i^{p_i}$ *is a GALF for* (1).

(β) *For every* $\mu \in \mathcal{M}_\Phi$ *there exists an* $i \in \{1, 2, \dots, n\}$ *with* $\int_Y f_i \, d\mu > 0$.

(γ) *There are* $p_1, p_2, \dots, p_n > 0$ *such that* $\sum_{i=1}^n p_i \int_Y f_i \, d\mu > 0$ *holds for every ergodic* $\mu \in \mathcal{M}_\Phi^E$.

Proof. We may assume by homogeneity that $p = (p_1, p_2, \dots, p_n) \in X$. Applying the lemmas for $h = \sum_i p_i f_i$, we obtain that

$$(\alpha) \Leftrightarrow \max_{p \in X} \min_{\mu \in \mathcal{M}_\Phi} \sum_{i=1}^n p_i \int_Y f_i \, d\mu > 0. \tag{33}$$

On the other hand, it is elementary to check that

$$(\beta) \Leftrightarrow \min_{\mu \in \mathcal{M}_\Phi} \max_{p \in X} \sum_{i=1}^n p_i \int_Y f_i \, d\mu > 0. \tag{34}$$

With $C = \mathcal{M}_\Phi$, $D = X$, $A = C_w^*(Y, R)$, $B = \mathbb{R}^n$, and $\Gamma(p, \mu) = \sum_i p_i \int_Y f_i \, d\mu$, the minimax theorem implies the equivalence of (α) and (β). Since the minimum in (33) is attained at an ergodic measure, the equivalence of (α) and (γ) follows. ∎

Example 4.5. Returning to the rock-scissors-paper game (4)–(5), note that

$$\mathcal{M}_\Phi = \left\{ \sum_{k=1}^3 q_k \delta_k \;\middle|\; q_k \geq 0, \sum_{k=1}^3 q_k = 1 \right\},$$

where δ_k is the Dirac measure at the kth vertex of the two-dimensional simplex, $k = 1, 2, 3$. By using homogeneity, condition (β) then translates into the requirement that

$$\text{for any } \; q > 0 \; \text{ there exists an } \; i \in \{1, 2, 3\} \text{ with } \; (Aq)_i > 0. \tag{35}$$

The equivalence of (7) and (35), for arbitrary $n \times m$ real matrices, is the well-known Farkas lemma on linear inequalities. Note that in an alternative proof of Theorem 4.4, the minimax theorem can be replaced by using an infinite-dimensional version of the Farkas lemma.

The integrals $\int_Y f_i \, d\mu$ are Liapunov exponents of μ. If μ is ergodic, then there exists a unique nonempty supporting subset $I \subset \{1, 2, \dots, n\}$ such that $\mu(X_I) = 1$ for the (relatively) open face $X_I := \{x \in X : x_i > 0 \text{ for } i \in I \text{ and } x_j = 0 \text{ for } j \notin I\}$. According to Lemma 5.1 in [57], $\int_Y f_i \, d\mu = 0$ for $i \in I$ (compare also Remark 5.4). The integrals $\int_Y f_i \, d\mu$ for $i \notin I$ are called *external Liapunov exponents*. Biologically, they describe the invasion rate of the missing species i at μ. For point measures $\delta_{\bar{x}}$, the external Liapunov exponents reduce to the external eigenvalues $f_i(\bar{x})$ at the boundary equilibrium $\bar{x}$; see [35]. For periodic orbits in Y, the external Liapunov exponents coincide with the (normalized) external Floquet exponents; see [57].

Combining Theorems 2.2 and 4.4, we see that (β) is a sufficient condition for robust permanence. This is a version of the main result in [57]. Strengthening condition (β) leads to the following result on "totally permanent systems" due to Mierczyński and Schreiber [52].

COROLLARY 4.6. *If for every $\mu \in \mathcal{M}_\Phi^E$ all external Liapunov exponents $\int_Y f_i d\mu$ are positive, then* (1) *and each of its subsystems are robustly permanent.*

Proof. This follows immediately from Theorems 2.2 and 4.4, together with the aforementioned Lemma 5.1 of [57] or Remark 5.4 below. Note that $P(x) = \prod_{i=1}^n x_i^{p_i}$ is a GALF for (1) for any choice of the exponents $p_i > 0$. ∎

Using Pesin theory, a converse result can also be shown; see [52].

5. Local GALFs and Morse decompositions. Throughout this section, let K be a nonempty Φ-invariant compact subset of Y, and let U be an open neighborhood of K in $\mathbb{R}^n_+$.

DEFINITION 5.1. *A continuous mapping $P_K : U \to \mathbb{R}$ is a GALF* for (1) on K *if*

(a)$_K$ $P_K(x) = 0$ *for all* $x \in U \cap \partial\mathbb{R}^n_+$, $P_K(x) > 0$ *for all* $x \in U \cap \operatorname{int}\mathbb{R}^n_+$;

(b)$_K$ P_K *is differentiable on* $U \cap \operatorname{int}\mathbb{R}^n_+$ *and* $p_i(x) := \frac{x_i}{P_K(x)} \frac{\partial P_K}{\partial x_i}$ *can be extended to a continuous function on* U *for every* i;

(c)$_K$ *for every* $y \in K$ *there is a positive constant* T_y *with the property that* $\int_0^{T_y} \sum_i p_i(\Phi(t,y)) f_i(\Phi(t,y))\, dt > 0$.

THEOREM 5.2. *If P_K is a GALF for* (1) *on K, then there exist an open neighborhood $\mathcal{N}_K$ of K in X and positive constants $\delta, \kappa_1, \kappa_2$ such that for each δ-perturbation*

$$P_K(\Psi(t,x)) \le \kappa_1 e^{\kappa_2 t} P_K(x) \quad \text{whenever} \quad \{\Psi(\tau,x) \mid t \le \tau \le 0\} \subset \mathcal{N}_K.$$

In particular, $\mathcal{N}_K \setminus Y$ does not contain entire trajectories of Ψ and, for each $x \in \mathcal{N}_K \setminus Y$, inclusion $\emptyset \ne \omega_\Psi(x) \subset K$ is impossible.

Proof. Reconsidering the respective proofs in sections 2 and 3, we see that the existence of a local GALF implies that both (17) and (11) remain valid in the local setting. ∎

The collection of Φ-invariant Borel probability measures on K is denoted by $\mathcal{M}_\Phi(K)$. Clearly $\mathcal{M}_\Phi(Y) = \mathcal{M}_\Phi$ and, for a general K, $\mathcal{M}_\Phi(K)$ can be identified with $\{\mu \in \mathcal{M}_\Phi : \mu(K) = 1\}$. The collection of ergodic measures in $\mathcal{M}_\Phi(K)$ is denoted by $\mathcal{M}_\Phi^E(K)$.

THEOREM 5.3. *The following properties are pairwise equivalent:*

$(\alpha)_K$ *There are $p_1, p_2, \dots, p_n > 0$ such that $P(x) = \prod_{i=1}^n x_i^{p_i}$ is a GALF for* (1) *on K.*

$(\beta)_K$ *For every $\mu \in \mathcal{M}_\Phi(K)$ there exists an $i \in \{1, 2, \dots, n\}$ with $\int_K f_i\, d\mu > 0$.*

$(\gamma)_K$ *There are $p_1, p_2, \dots, p_n > 0$ such that $\sum_{i=1}^n p_i \int_K f_i d\mu > 0$ for all $\mu \in \mathcal{M}_\Phi^E(K)$.*

Proof. This is the localized version of Theorem 4.4, replacing Y by K, with the same proof. ∎

Remark 5.4. Assume that $K \subset \{y \in Y : y_n > 0\}$. Then

$$\int_K f_n\, d\mu = 0 \quad \text{for each } \mu \in \mathcal{M}_\Phi(K).$$

In fact, a twofold application of the Φ-invariance of μ implies via Fubini's theorem that

$$\begin{aligned}\int_K f_n\, d\mu &= \int_0^1 \int_K f_n(\Phi(t,\cdot))\, d\mu dt = \int_K \int_0^1 f_n(\Phi(t,\cdot))\, dt d\mu \\ &= \int_K \{\log(\Phi_n(t,\cdot))\}_{t=0}^{t=1}\, d\mu = \int_K \log(\Phi_n(1,\cdot))\, d\mu - \int_K \log(\Phi_n(0,\cdot))\, d\mu = 0.\end{aligned}$$

The property established above (proved differently in [57, Lem. 5.1], using the ergodic theorem and Poincaré's recurrence theorem) helps check whether $(\beta)_K$ is satisfied or not.

Schreiber [57] (working on $\mathbb{R}^n_+$) defines an invariant probability measure $\mu \in \mathcal{M}_\Phi$ to be *unsaturated* if $\max_{1\le i\le n} \int_Y f_i \, d\mu > 0$, i.e., at least one external Liapunov exponent is positive. (For point measures this reduces to the notion of an unsaturated equilibrium from [35].) He calls a compact invariant set $K \subset Y$ unsaturated if every $\mu \in \mathcal{M}_\Phi(K)$ is unsaturated. By our Theorem 5.3, K is unsaturated if and only if there exists a local GALF near K of the standard form $\prod_i x_i^{p_i}$. One of the main results in [57] and [26] says that if Y has a Morse decomposition with all Morse sets being unsaturated, then (1) is robustly permanent. This result is generalized as follows.

THEOREM 5.5. *Let $M_1, M_2, \dots, M_\ell$ be a Morse decomposition on Y for $\Phi|_Y$. Further, for $k = 1, 2, \dots, \ell$, let U_k be an open neighborhood of M_k in Y and let $P_k : U_k \to \mathbb{R}$ be a GALF for* (1) *on M_k. Then* (1) *is robustly permanent.*

Proof. Arguing as in the first paragraph of the proof of Theorem 2.2, we obtain that there are three constants $c, \delta, T > 0$ and, for $k = 1, 2, \dots, \ell$, there is an open neighborhood N_k of M_k in U_k with $\mathrm{cl}(N_j) \cap \mathrm{cl}(N_k) = \emptyset$ for $j \ne k$ and the property as follows. Given $x \in N_k \setminus Y$, $k = 1, 2, \dots, \ell$, arbitrarily, there exists a time $T_x \in (0, T]$ such that

$$P_k(\Psi(T_x, x)) > (1 + c)P_k(x) \quad \text{for every } \delta\text{-perturbation (2) of (1).}$$

We claim that, for δ sufficiently small,

$$\gamma_\Psi(x) \subset \mathcal{B}[Y, \delta] \Rightarrow \alpha_\Psi(x) \cup \omega_\Psi(x) \subset \bigcup_{k=1}^{\ell} N_k. \tag{36}$$

Since $\bigcup_k M_k$ is the intersection of a finite collection of attractor–repeller pairs, there is no loss of generality in assuming that $\ell = 2$ and that (M_1, M_2) is an attractor–repeller pair for $\Phi|_Y$.

Since M_1 is an attractor for $\Phi|_Y$, there exists a compact neighborhood S_1 of M_1 in N_1 satisfying $\Phi(\mathbb{R}_+, Y \cap S_1) \subset N_1$. We point out next that, for δ sufficiently small,

$$\gamma_\Psi(z) \subset \mathcal{B}[Y, \delta] \text{ plus } z \in S_1 \Rightarrow \gamma^+_\Psi(z) \subset N_1. \tag{37}$$

To the contrary, suppose that, for each $j = 1, 2, \dots$, there exists a $\frac{1}{n}$-perturbation (2) of (1), a $z_j \in S_1$, and a time $t_j > 0$ satisfying $\gamma_{\Psi_j}(z_j) \subset \mathcal{B}[Y, \frac{1}{j}]$ but $w_j = \Psi_j(t_j, z_j) \notin N_1$. We may assume that $z_j \in \partial S_1$, $w_j \in \partial N_1$, $\Psi_j((0, t_j), z_j) \subset N_1 \setminus S_1$ and, by compactness, $z_n \to z_0$ and $w_n \to w_0$ for some $z_0 \in Y \cap \partial S_1$ and $w_0 \in Y \cap \partial N_1$. We distinguish two cases according to whether $\{t_j\} \subset \mathbb{R}_+$ is bounded or not. By passing to a subsequence, we may assume that $t_j \to t_0$ for some $t_0 \in \mathbb{R}_+$ or $t_j \to \infty$. If $t_j \to t_0$, then $\Phi(t_0, z_0) = w_0$, a contradiction. If $t_j \to \infty$, we may assume that $q_j = \Psi_j(t_j/2, z_j) \to q$ for some $q \in Y \cap \mathrm{cl}(N_1 \setminus S_1)$. It is readily checked that $\gamma_\Phi(q) \subset Y \cap \mathrm{cl}(N_1 \setminus S_1)$, a contradiction.

By continuity (and passing to a smaller δ if necessary), we see there exist positive times T_1, $T_2 > T_1$ such that

$$\Psi([T_1, T_2], \mathcal{B}[Y \setminus (N_1 \cup N_2), \delta]) \subset S_1 \quad \text{for every } \delta\text{-perturbation (2) of (1).}$$

In view of property (37), this ends the proof (of case $\ell = 2$) of (36).

The rest is easy. For each $k = 1, 2, \dots, \ell$, the Zubov–Ura–Kimura argument we used in the last two paragraphs of the proof of Theorem 2.2 applies in $\mathrm{cl}(N_k)$ individually. ∎

Remark 5.6. The proof of Theorem 5.5 shows that any continuous function $P : X \to \mathbb{R}$ satisfying conditions (a) and $P|_{N_k} = P_k|_{N_k}$, $k = 1, 2, \dots, \ell$, satisfies condition (9), too. Thus, in a technical sense, local GALFs can be joined together to a global "nearly-GALF."

6. Discrete-time analogues. Mutatis mutandis, all the previous results remain valid for discrete-time dynamical systems.

With $F_i : X \to \mathbb{R}$, $i = 1, 2, \dots, n$, continuous, consider a mapping of the form

$$\mathcal{F} : X \to X, \quad x \to (x_1 F_1(x), x_2 F_2(x), \dots, x_n F_n(x)). \tag{38}$$

Throughout this section, it is assumed that $\mathcal{F}$ is a self-homeomorphism of X. Brouwer's open mapping theorem implies that $\mathcal{F}(X \setminus Y) = X \setminus Y$ and $\mathcal{F}(Y) = Y$. In particular, $F_i(x) > 0$ for each $x \in X \setminus Y$, $i = 1, 2, \dots, n$. Throughout this section, we assume further that $F_i(y) > 0$ for each $y \in Y$, $i = 1, 2, \dots, n$.

Our next result implies that this latter assumption is quite natural. It will also be crucial in establishing Lemma 7.2, the starting point of the theory of discretizations of Kolmogorov type in the next chapter, and it was used already in the proof of Theorem 3.4.

Surjectivity theorem. *Let $F_i : X \to \mathbb{R}_+$, $i = 1, 2, \dots, n$, be continuous functions, $\sum_i x_i F_i(x) = 1$ for each $x \in X$, and consider the mapping $\mathcal{F} : X \to X$, $x \to (x_1 F_1(x), x_2 F_2(x), \dots, x_n F_n(x))$. Then $\mathcal{F}(S) = S$ for each subsimplex S of X.*

Proof. By the particular form of our mapping, this is certainly true for the zero-dimensional subsimplices (vertices) of X. For a k-member subset $\{i_1, i_2, \dots, i_k\}$ of $\{1, 2, \dots, n\}$, consider the subsimplex of the form $S = \{x \in X : x_{i_1} = x_{i_2} = \cdots = x_{i_k} = 0\}$. Applying $\mathcal{F}_i(x) = x_i F_i(x)$ for $i = i_1, i_2, \dots, i_k$, we obtain that $\mathcal{F}(S) \subset S$. By induction on the subsimplices, we may assume that $S = X$ and $\mathcal{F}(s) = s$ for each facet s of $S = X$. Consider a point $p_0 \in \mathrm{int}(S)$ arbitrarily chosen. For any $\lambda \in [0, 1]$, any facet s of S, and any point $p \in s$, the convexity of s implies that $(1 - \lambda)p + \lambda\mathcal{F}(p) \in s$. It follows that

$$(1 - \lambda)p + \lambda\mathcal{F}(p) \neq p_0 \quad \text{whenever } p \in \partial S \text{ and } \lambda \in [0, 1].$$

By the homotopy property of Brouwer's degree, it follows that

$$\deg(\mathcal{F}, p_0, \mathrm{int}(S)) = \deg(\mathrm{id}_{R^n}, p_0, \mathrm{int}(S)),$$

where id_{R^n} denotes the identity on $\mathbb{R}^n$. Since $\deg(\mathrm{id}_{R^n}, p_0, \mathrm{int}(S)) = 1$, the existence property of the degree implies that $p_0 \in \mathcal{F}(S)$. ∎

Corollary 6.1. *In addition, assume that F_i, $i = 1, 2, \dots, n$, is of class C^1 (in the sense that F_i admits a C^1 extension $\hat{F}_i : U_i \to \mathbb{R}$ defined on an open neighborhood U_i of X in $\mathbb{R}^n$) and that $\mathcal{F}$ is a C^1 self-diffeomorphism of X. Then $F_i(y) > 0$ for each $y \in Y$, $i = 1, 2, \dots, n$.*

Proof. Pick $y \in Y$ arbitrarily. For index j satisfying $y_j \neq 0$, inequality $F_j(y) > 0$ is a direct consequence of the surjectivity theorem when applied to the X-facet $S_j = \{x \in X : x_j = 0\}$. For j satisfying $y_j = 0$, inequality $F_j(y) > 0$ follows from the diffeomorphism assumption. To the contrary, assume that $F_j(y) = 0$ (and $y_j = 0$). A direct computation shows that the jth row of the Jacobian of $\mathcal{F}$ evaluated at y equals $(0, 0, \dots, 0)$, a contradiction. ∎

The discrete-time version of an ALF for $\mathcal{F}$ [40] is a continuous mapping $R : X \to \mathbb{R}$ with the following properties:

(d) $R(x) = 0$ for all $x \in Y$, $R(x) > 0$ for all $x \in X \setminus Y$.

(e) There exists a continuous function $r : X \to \mathbb{R}$ such that $r(x) = \log(R(\mathcal{F}(x))) - \log(R(x))$ whenever $x \in X \setminus Y$.

(f) For every $y \in Y$ there is a positive integer $N_y > 0$ with the property that $\sum_{k=1}^{N_y} r(\mathcal{F}^{k-1}(y)) > 0$.

In contrast to the continuous-time case, we did not find a reasonable analogue of the notion of GALF for discrete time. Hence for studying robust permanence in discrete-time systems, we restrict ourselves to the standard ALF of the form $R(x) = \prod_{i=1}^{n} x_i^{r_i}$ with $r_i > 0$, $i = 1, 2, \ldots, n$. For this choice of R, condition (e) holds with $r(x) = \sum_i r_i \log F_i(x)$.

Remark 6.2. One can characterize this R by a functional equation. More precisely, if the continuous mappings $r_i, R : X \to \mathbb{R}$ satisfy condition (d) and $\log R(\mathcal{F}(x)) - \log R(x) = \sum_i r_i(x) \log F_i(x)$ for arbitrary $\mathcal{F}$, then $r_i(x) = r_i$ and $R(x) = c_n \prod_{i=1}^{n} x_i^{r_i}$ for some positive constants r_i, $i = 1, 2, \ldots, n$, and c_n. For a proof, see [19].

With $G_i : X \to \mathbb{R}$, $i = 1, 2, \ldots, n$, continuous, consider δ-perturbations of $\mathcal{F}$ of the form $\mathcal{G} : X \to X$, $x \to (x_1 G_1(x), x_2 G_2(x), \ldots, x_n G_n(x))$, where $|G_i(x) - F_i(x)| < \delta$, $i = 1, 2, \ldots, n$. It is of course assumed that $\sum x_i G_i(x) = 1$ for each $x \in X$. We assume further that $\mathcal{G}$ is a self-homeomorphism of X. If $\mathcal{G}$ is permanent, then $(\mathcal{A}_{\mathcal{G}}, Y)$ forms an attractor–repeller pair, where $\mathcal{A}_{\mathcal{G}}$ denotes the maximal compact $\mathcal{G}$-invariant set in $X \setminus Y$. In analogy to the relation between (2) and (1), we say that $\mathcal{G}$ is a δ-perturbation of $\mathcal{F}$ if $|G_i(x) - F_i(x)| < \delta$ for each $x \in X$ and $i = 1, 2, \ldots, n$.

THEOREM 6.3. *If there is an ALF for $\mathcal{F}$, then $\mathcal{F}$ is permanent. Moreover, assume that for some constants $r_i > 0$ suitably chosen, $R(x) = \prod_{i=1}^{n} x_i^{r_i}$ is an ALF for $\mathcal{F}$. Then $\mathcal{F}$ is robustly permanent. There are a $\delta > 0$ and a compact subset S of $X \setminus Y$ with the properties as follows. Every δ-perturbation $\mathcal{G}$ of $\mathcal{F}$ is permanent and $\mathcal{A}_{\mathcal{G}}$ is contained in S.*

Proof. The proof of Theorem 2.2 can be repeated. The computations are based on the formulae

$$\log(R(\mathcal{F}^{N_x}(x))) - \log(R(x)) = \sum_{k=1}^{N_x} r(\mathcal{F}^{k-1}(x))$$

and

$$\log(R(\mathcal{G}^{N_x}(x))) - \log(R(x)) = \sum_{k=1}^{N_x} \sum_{i=1}^{n} r_i \cdot \log(G_i(\mathcal{G}^{k-1}(x))),$$

respectively. The last step is the application of the discrete-time version of the Zubov–Ura–Kimura theorem. ∎

The first statement of Theorem 6.3 is due to [40]; see also [36]. The robustness result is new.

The set of $\mathcal{F}$-invariant Borel probability measures on Y is denoted by $\mathcal{M}_{\mathcal{F}}$. When combined with Theorem 6.3, our next result establishes a sufficient condition for robust permanence of $\mathcal{F}$. Note that inequality $\int_Y \log(F_i)\, d\nu > 0$ is stronger than the (seemingly) "more natural" inequality $\int_Y F_i\, d\nu > 1$. A special case of heteroclinic cycles was treated in [24].

THEOREM 6.4. *The following properties are equivalent:*

$(\alpha)^d$ *For some $r_1, r_2, \ldots, r_n > 0$ suitably chosen, $R(x) = \prod_{i=1}^{n} x_i^{r_i}$ is an ALF for $\mathcal{F}$.*

$(\beta)^d$ *For every* $\nu \in \mathcal{M}_{\mathcal{F}}$ *there exists an* $i \in \{1, 2, \dots, n\}$ *with* $\int_Y \log F_i \, d\nu > 0$.

$(\gamma)^d$ *There are* $r_1, r_2, \dots, r_n > 0$ *such that* $\sum_{i=1}^n r_i \int_Y \log F_i \, d\nu > 0$ *holds for every ergodic* $\nu \in \mathcal{M}_{\mathcal{F}}^E$.

Proof. The proof of Theorem 4.4 can be repeated. Almost no changes are needed. The method of proving Lemma 4.1 yields for each $h \in C(Y, \mathbb{R})$ that

$$\min_{\nu \in \mathcal{M}_{\mathcal{F}}} \int_Y h \, d\nu = \min_{y \in Y} \limsup_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} h(\mathcal{F}^{k-1}(y)).$$

The compactness argument we used in proving Lemma 4.2 implies that inequality $\min_{\nu \in \mathcal{M}_{\mathcal{F}}} \int_Y h \, d\nu > 0$ is equivalent to the following assertion: For every $y \in Y$ there is a positive integer $N_y > 0$ such that $\sum_{k=1}^{N_y} h(\mathcal{F}^{k-1}(y)) > 0$. We may assume by homogeneity that $r = (r_1, r_2, \dots, r_n) \in X$. With $h = \sum r_i \cdot \log(F_i)$, the minimax theorem applies. □

THEOREM 6.5. *Let* $M_1, M_2, \dots, M_\ell$ *be a Morse decomposition on* Y *for* $\mathcal{F}|_Y$. *Further, for* $k = 1, 2, \dots, \ell$, *let* U_k *be an open neighborhood of* M_k *in* Y *and let* $R_k : U_k \to \mathbb{R}$ *be an ALF for* $\mathcal{F}$ *on* M_k. *Then* $\mathcal{F}$ *is permanent. Moreover, assume that each* R_k *is of the special form* $R_k(x) = \prod_{i=1}^n x_i^{r_i^k}$ *for some positive constants* $r_1^k, r_2^k, \dots, r_n^k$. *Then* $\mathcal{F}$ *is robustly permanent.*

Proof. The method we used in proving Theorem 5.5 applies. □

The formulation of the discrete-time version of Theorems 3.1, 3.4, 5.2, and 5.3 is left to the reader.

7. Discretizations of Kolmogorov type. We discuss definition and basic properties of $\mathcal{P}$th order one-step discretizations of (1).

Let h_0 be a positive constant. Let $\mathcal{P} \geq 1$, $k \geq 0$ be integers with $\mathcal{P} + k \geq 2$. Assume that $f_1, f_2, \dots, f_n$ are $C^{\mathcal{P}+k+1}$ functions. The $C^{\mathcal{P}+k+1}$ property on closed sets like X (or $[0, h_0] \times X$) is understood as the existence of a $C^{\mathcal{P}+k+1}$ extension defined on an open neighborhood of X in $\mathbb{R}^n$ (or of $[0, h_0] \times X$ in $\mathbb{R} \times \mathbb{R}^n$). Consider a $C^{\mathcal{P}+k+1}$ discretization operator $\varphi : [0, h_0] \times X \to \mathbb{R}^n$. We assume that φ is of order $\mathcal{P}$, i.e., there exists a positive constant K (depending only on $\{f_i\}_{i=1}^n$) such that

$$|\Phi(h, x) - \varphi(h, x)| \leq K h^{\mathcal{P}+1} \quad \text{for all } h \in [0, h_0] \text{ and } x \in X.$$

We require also that φ is locally determined by $\{f_i\}_{i=1}^n$; i.e., we assume the existence of a continuous function $\Delta : [0, h_0] \to [0, \infty)$ such that $\Delta(0) = 0$ and, for all $h \in (0, h_0]$ and $x \in X$, $\varphi(h, x)$ is determined solely by the restriction of $\{f_i\}_{i=1}^n$ to $\mathcal{B}(x, \Delta(h))$. All these assumptions are satisfied if φ comes from a (general r-stage explicit or implicit) Runge–Kutta method. The standard theory of discretization operators (see, e.g., Stuart and Humphries [60]) implies that for all h sufficiently small, say $h \in [0, h_0]$, $\varphi(h, \cdot)$ is a $C^{\mathcal{P}+k+1}$ diffeomorphism of X onto $\varphi(h, X)$.

Now we are in a position to define discretizations of Kolmogorov type. Besides the above requirements on differentiability, consistency, and determinacy (these three were grouped together in [7] for the first time), two further conditions on a general discretization operator are imposed.

DEFINITION 7.1. *We say that our discretization operator is* of Kolmogorov type *on* X *for* (1) *if, for each* $i = 1, 2, \dots, n$, *there exists a* $C^{\mathcal{P}+k+1}$ *function* $q_i : [0, h_0] \times X \to \mathbb{R}$ *satisfying*

$$\varphi_i(h, x) = x_i q_i(h, x) \quad \text{whenever } h \in [0, h_0] \text{ and } x \in X \tag{39}$$

and, in addition, $\varphi(h, X) \subset X$ *for each* $h \in [0, h_0]$.

LEMMA 7.2. *Let φ be a discretization operator of Kolmogorov type on X for* (1). *Then, for all h sufficiently small, say $h \in [0, h_0]$, $\varphi(h, \cdot)$ defines a $C^{\mathcal{P}+k+1}$ discrete-time dynamical system on X.*

Proof. We know already that $\varphi(h, \cdot)$ is a $C^{\mathcal{P}+k+1}$ diffeomorphism of X onto $\varphi(h, X) \subset X$. The surjectivity theorem applies. ∎

Remark 7.3. In accordance with (39), the solution operator of (1) satisfies

$$\Phi_i(h, x) = x_i Q_i(h, x) \quad \text{whenever } h \in [0, h_0] \text{ and } x \in X, \tag{40}$$

where $Q : [0, h_0] \times X \to \mathbb{R}^n$ is a $C^{\mathcal{P}+k}$ function defined by

$$Q_i(h, x) = \int_0^1 \frac{d}{dx_i} \Phi_i(h, x_1, \ldots, x_{i-1}, \theta x_i, x_{i+1}, \ldots, x_n)\, d\theta, \quad i = 1, 2, \ldots, n.$$

Actually, Q is of class $C^{\mathcal{P}+k+1}$. Existence and continuity of the last derivative is a consequence of the $C^{\mathcal{P}+k+1}$ parametrized version of the Picard–Lindelöf theorem. In fact, with $x \in X$ as a parameter, let $z(\cdot; x)$ denote the solution of the initial value problem

$$\dot{z}_i = z_i f_i(x_1 z_1, x_2 z_2, \ldots, x_n z_n) \quad \text{and} \quad z_i(0) = 1, \qquad i = 1, 2, \ldots, n.$$

Since $(x_1 z_1(\cdot; x), x_2 z_2(\cdot; x), \ldots, x_n z_n(\cdot; x))$ is a solution to (1), we have by uniqueness that $z(t; x) = Q(t, x)$ for all $t \in R$ and $x \in X$.

Example 7.4. Let $\vartheta : [0, h_0] \times X \to \mathbb{R}^n$ be a discretization operator coming from a (general r-stage explicit or implicit) Runge–Kutta method. It is a straightforward but rather lengthy task to check that, for all h sufficiently small, say $h \in [0, h_0]$, formula

$$\varphi_i(h, x) = \frac{\vartheta_i(h, x)}{\sum_j \vartheta_j(h, x)}\,, \qquad x \in X, \quad i = 1, 2, \ldots, n,$$

makes sense and defines a $\mathcal{P}$th order discretization operator of Kolmogorov type on X for (1). For example, the explicit Euler method leads to

$$\varphi_i^E(h, x) = x_i \frac{1 + h f_i(x)}{1 + h \sum_j x_j f_j(x)}\,, \quad (h, x) \in [0, h_0] \times X, \qquad i = 1, 2, \ldots, n\,,$$

a first order discretization operator of Kolmogorov type.

The difference between exact and discretized solutions of (1) on finite-time intervals can be estimated as follows.

LEMMA 7.5. *Let φ be a $\mathcal{P}$th order discretization operator of Kolmogorov type on X for* (1). *Given $T > 0$ arbitrarily, there exists a positive constant $\kappa(T)$ such that for any $M = 0, 1, 2, \ldots$ with $Mh \le T$, the estimate*

$$|\Phi_i(Mh, x) - \{\varphi^M(h, \cdot)\}_i(x)| \le x_i \cdot \kappa(T) \cdot h^{\mathcal{P}}, \quad (h, x) \in [0, h_0] \times X \tag{41}$$

holds true. (Here of course $\{\varphi^M(h, \cdot)\}_i$ denotes the ith coordinate function of the Mth iterate of the discretization mapping $\varphi(h, \cdot)$, $i = 1, 2, \ldots, n$.)

Proof. Writing out the coordinate functions explicitly, we find that methods of deriving the standard error estimate $|\Phi(Mh, x) - \varphi^M(h, x)| \le \kappa_0(T) h^{\mathcal{P}}$ (e.g., in [60]) apply and κ is an exponential function of T. For details, see [21]. ∎

LEMMA 7.6. *The previous lemma holds true for variable stepsize sequences. More precisely, given $T > 0$ arbitrarily and $\kappa(T)$ denoting the same constant as in* (41)*, the estimate*

$$\left|\Phi_i\left(\sum_{m=1}^{M} h_m, x\right) - \{\varphi(h_M,\cdot)\circ\cdots\circ\varphi(h_1,\cdot)\}_i(x)\right| \le x_i\cdot\kappa(T)\cdot\left(\max_{1\le m\le M} h_m\right)^{\mathcal{P}}$$

holds true whenever $h_m \in (0, h_0]$, $m = 1, 2, \ldots, M$, with $\sum_m h_m < T$, and $x \in X$.

Proof. The proof is almost the same as that for constant stepsizes. ∎

8. Permanence for discretizations. Results in this section fit well in the list of papers in [60] on attraction, Liapunov functions, and discretization. They are particularly closely related to continuity results on exponentially attracting attractors in [4] and on convergence rates of perturbed attracting sets with vanishing perturbation [25]. For the qualitative theory of discretizations in general, see the monograph [60] as well as the fundamental paper [48].

LEMMA 8.1. *Fix $p_i > 0$, $i = 1, 2, \ldots, n$, and consider mapping $P : X \to \mathbb{R}_+$, $P(x) = \prod_{i=1}^n x_i^{p_i}$. Then P is an ALF for* (1) *if and only if P is an ALF for $\mathcal{F} = \Phi(1,\cdot)$.*

Proof. By letting $b = 1$ and $a = k - 1$, $k = 1, 2, \ldots, N$, in the identity $\log(Q_i(b, \Phi(a, y))) = \int_a^{a+b} f_i(\Phi(t, y))\, dt$ (a simple consequence of (40)) and forming the respective linear combinations,

$$\sum_{k=1}^{N}\sum_{i=1}^{n} p_i\cdot\log(Q_i(1,(\mathcal{F}^{k-1}(y)))) = \int_0^N \sum_{i=1}^{n} p_i f_i(\Phi(t,y))\, dt$$

holds for each $y \in Y$ and $N = 1, 2, \ldots$. Consequently, if P is an ALF for $\mathcal{F}$, then P is an ALF for (1) and $T_y = N_y$. Conversely, assume that $\int_0^{T_y}\sum f_i(\Phi(t,y))\, dt > 0$ for some $T_y > 0$. The compactness argument we used in proving Lemma 4.2 implies that $\int_0^{\tau_y}\sum f_i(\Phi(t,y))\, dt > 0$ for some positive integer τ_y. Thus P is an ALF for $\mathcal{F}$ and $N_y = \tau_y$. ∎

Remark 8.2. Together with Lemma 4.1, a similar argument implies that

$$\min_{\nu\in\mathcal{M}_{\Phi(1,\cdot)}}\int_Y \log(Q_i(1,\cdot))\, d\nu = \min_{\mu\in\mathcal{M}_\Phi}\int_Y f_i\, d\mu \qquad \text{for each } \ i = 1, 2, \ldots, n.$$

This is somewhat strange because $\mathcal{M}_\Phi \subset \mathcal{M}_{\Phi(1,\cdot)}$ and the set $\mathcal{M}_{\Phi(1,\cdot)}$ is usually a much larger subset of $\mathcal{M}$ than $\mathcal{M}_\Phi$. It is not hard to establish that the dependence of $\mathcal{M}_{\Phi(t,\cdot)}$ on the parameter $t \in (0,\infty)$ is weakly-$*$ upper semicontinuous. Moreover, if $\mathcal{U}$ is an open neighborhood of $\mathcal{M}_\Phi$ in the weak-$*$ topology of $\mathcal{M}$, then $\mathcal{M}_{\Phi(t,\cdot)} \subset \mathcal{U}$ for $|t|$ sufficiently small. Similarly, if φ is a discretization operator of Kolmogorov type, then there exists a positive constant $h_{\mathcal{U}}$ such that $\mathcal{M}_{\varphi(h,\cdot)} \subset \mathcal{U}$ whenever $0 < h \le h_{\mathcal{U}}$. (A detailed proof of this latter statement is contained in [21].) No upper semicontinuity result holds true for (the closure of the union of) supports of (all) invariant measures. On the general problem of measures and discretization, we recommend [14] and the references therein. Several upper semicontinuity results of numerical dynamics are contained also in [60].

Lemma 8.1 enables us to give a short proof for permanence under discretization.

THEOREM 8.3. *Assume that $P(x) = \prod_{i=1}^n x_i^{p_i}$ is an ALF for* (1)*. Let φ be a $\mathcal{P}$th order discretization operator of Kolmogorov type for* (1)*. Then, for all h sufficiently*

small, say $h \in (0, h_0]$, Y is a repeller for the discrete-time dynamical system induced by $\varphi(h,\cdot)$. In addition, there is a compact subset S of $X \setminus Y$ with the property that the dual attractor $\mathcal{A}_{\varphi(h,\cdot)}$ is contained in S, $h \in (0, h_0]$.

Proof. Assume that $0 < h \le h_0 < 1$ and consider the positive integer M_h satisfying $M_h h < 1 \le (M_h + 1)h$. Since[6] Y is $\varphi(h,\cdot)$-invariant, Y is a repeller for $\varphi(h,\cdot)$ if and only if Y is a repeller for $\varphi^{M_h}(h,\cdot)$. Combining Lemma 8.1 and Theorem 6.3, we see it is enough to point out that $\mathcal{G} = \varphi^{M_h}(h,\cdot)$ is a δ-perturbation of $\mathcal{F} = \Phi(1,\cdot)$.

In fact, for each $x \in X$, $h \in (0, h_0]$ and $i = 1, 2, \ldots, n$, inequality $0 < 1 - M_h h \le h$ implies that

$$|\Phi_i(1,x) - \Phi_i(M_h h, x)| = |\Phi_i(1 - M_h h, \Phi(M_h h, x)) - \Phi_i(M_h h, x)| \le x_i \cdot (e^{Kh} - 1)$$

with $K = \max_{1\le i\le n} \max_{x\in X} |f_i(x)|$. We conclude via (41) that

$$|\Phi_i(1,x) - \{\varphi^{M_h}(h,\cdot)\}_i(x)| \le x_i \left\{(e^{Kh} - 1) + \kappa(1) \cdot h^{\mathcal{P}}\right\}. \tag{42}$$

Note that the coefficient of x_i on the right-hand side of (42) approaches zero as $h \to 0^+$. Thus $\mathcal{G} = \varphi^{M_h}(h,\cdot)$ is a δ-perturbation of $\mathcal{F} = \Phi(1,\cdot)$ for h small enough and Theorem 6.3 applies. ∎

We do not know if Theorem 8.3 is true for a general GALF. The main difficulty is in proving the inequality

$$|P(\Phi(Mh, x)) - P(\varphi^M(h,x))| \le \tilde{\kappa}_1(T) h^{\mathcal{P}} \cdot P(x) \quad \text{if } 0 < T,\ 0 \le Mh \le T, \tag{43}$$

which seems to be a rather delicate matter.

Remark 8.4. Assume that the conditions of Theorem 8.3 are all satisfied. Combining Theorem 4.4 and the discretization result in Remark 8.2, one can establish the existence of positive constants c_*, h_* with the following property: For every stepsize $h \in (0, h_*]$ and $\mu_h \in \mathcal{M}_{\varphi(h,\cdot)}$ there exists an $i \in \{1, 2, \ldots, n\}$ with $\int_Y \log q_i(h,\cdot)\, d\mu_h > c_* h$. An alternative presentation of a great part of sections 8, 9, 10, and 11 can be centered around (the local version of) this inequality.

9. Variable stepsize discretizations. In this section we present a generalization of Theorem 8.3 for variable stepsize discretizations.

The natural framework of handling variable stepsize sequences is that of nonautonomous dynamics. For general considerations, including several attractor definitions in the nonautonomous setting, we refer to [45]. In what follows we restrict ourselves to recalling the concept of cocycle attractors/repellers and to presenting a special case of the key result from Kloeden and Schmalfuss [44].

THEOREM 9.1. *Assume that* (1) *is permanent and let φ be a discretization operator of Kolmogorov type for* (1). *In addition, let $\mathcal{U}_Y$ be an open neighborhood of Y in X, let $\mathcal{U}_{\mathcal{A}_\Phi}$ be an open neighborhood of $\mathcal{A}_\Phi$ in X, and assume that $\mathcal{U}_Y \cap \mathcal{U}_{\mathcal{A}_\Phi} = \emptyset$. Then there are positive constants h_* and τ with the following properties: Given an arbitrary set C with $\mathcal{U}_{\mathcal{A}_\Phi} \subset C \subset X \setminus \mathcal{U}_Y$ and a doubly infinite stepsize sequence $\mathbf{h} = \{h_k\}_{k=-\infty}^{\infty}$ with $\sum_{k=1}^{\infty} h_k = \sum_{k=-\infty}^{0} h_k = \infty$ and $\|\mathbf{h}\| = \sup h_k \le h_*$,*

$$\{\varphi(h_M,\cdot) \circ \cdots \circ \varphi(h_1,\cdot)\}(C)\ ,\ \{\varphi(h_0,\cdot) \circ \cdots \circ \varphi(h_{-M},\cdot)\}(C) \subset \mathcal{U}_{\mathcal{A}_\Phi} \tag{44}$$

[6] Consider $W = [-1, 1] \subset \mathbb{R}$, $\mathcal{F}_0(w) = -w^3$ for $w \in W$. Then $\{1\}$ is a repeller for $\mathcal{F}_0^2$ but not for $\mathcal{F}_0$. The reason is that $\{1\}$ is not $\mathcal{F}_0$-invariant.

whenever $\sum_{k=1}^{M} h_k \geq \tau$ *and* $\sum_{k=-M}^{0} h_k \geq \tau$. *In addition, the set*

$$A(\mathbf{h}) = \bigcap\nolimits_{M\geq 0} \mathrm{cl}\left(\bigcup\nolimits_{m\geq M} \{\varphi(h_0,\cdot)\circ\cdots\circ\varphi(h_{-m},\cdot)\}(C)\right)$$

(45) *is independent of* C *and is contained in* $\mathrm{cl}(\mathcal{U}_{\mathcal{A}_\Phi})$,

(46) $$\mathrm{cl}\left(\bigcup\nolimits_{m\geq M} \{\varphi(h_0,\cdot)\circ\cdots\circ\varphi(h_{-m},\cdot)\}(C)\right) \to A(\mathbf{h})$$

in the Hausdorff metric as $M\to\infty$, *and, with* $\theta^m\mathbf{h}$ *denoting the doubly infinite shifted stepsize sequence defined by* $(\theta^m\mathbf{h})_k = h_{k+m}$,

(47) $$\{\varphi(h_m,\cdot)\circ\cdots\circ\varphi(h_1,\cdot)\}(A(\mathbf{h})) = A(\theta^m\mathbf{h}) \quad \textit{for each} \quad m=1,2,\ldots.$$

Proof. This is a restatement of Theorems 3.1 and 4.5 of [44] within the context of the present paper. Actually, the original results in Kloeden and Schmalfuss [44] are proved under the additional requirement

(48) $$\sup\{h_k/h_\ell \mid k,\ell\in\{0,\pm1,\pm2,\ldots\}\} \leq \text{const.}$$

The starting point of their proof is a classical result in converse Liapunov theory, Theorem 22.5 of Yoshizawa [67] on the existence of Lipschitz continuous Liapunov functions. However, when starting from Conley's C^∞ Liapunov function for the attractor–repeller pair $(\mathcal{A}_\Phi, Y)$, condition (48) turns out to be irrelevant. It is enough to replace Lemma 4.1 of [44] by Lemma 9.2 below and to reconsider the Kloeden–Schmalfuss argumentation. We find that Theorem 9.1 holds true for free stepsize sequences (subject only to the requirements $\sum_{k=1}^{\infty} h_k = \sum_{k=-\infty}^{0} h_k = \infty$ and $\|\mathbf{h}\| = \sup h_k \leq h_*$). ∎

For convenience, recall Lemma 1 of [20], which we "inserted" in the original proof of Theorem 9.1 in [44] above.

LEMMA 9.2. *There exists a* C^∞ *function* $V: X\to[0,1]$ *with the following properties: For every* $x\in X\setminus(\mathcal{A}_\Phi\cup Y)$, *function* $\mathbb{R}\to(0,1)$, $t\to\Phi(t,x)$ *is strictly decreasing, and, in addition,* $V^{-1}(0)=\mathcal{A}_\Phi$, $V^{-1}(1)=Y$. *Finally, for* $c\in(0,1)$ *arbitrarily given, there exists a positive constant* $h^*(c)$ *such that* $V(\varphi(h,x))<c$ *whenever* $h\in(0,h^*(c)]$ *and* $V(x)\leq c$.

Proof. This is a discretization consequence of Theorem 6.12 of Akin [2]. Details can be found in [20]. ∎

Most results in [44], [45] are stated and proved for abstract cocycles with the shift operator θ acting on a compact parameter space. Having applications to stochastic numerics in mind, no attempt is made in these papers to lift/weaken condition (48) in the simplest special case of deterministic discretizations with variable stepsize. The set $A(\mathbf{h})$ is called a *cocycle* or *pull-back attractor*. Properties (44), (45), (46), and (47) are called the *upper semicontinuity, uniqueness, pull-back convergence*, and *equivariance properties*, respectively. Elementary examples show that, together with the stepsize sequence $\mathbf{h}=\{h_k\}_{k=-\infty}^{\infty}$, the accompanying push-forward sequence of sets $\{\varphi(h_M,\cdot)\circ\cdots\circ\varphi(h_1,\cdot)\}(C)$ may also exhibit an oscillating behavior in $\mathcal{U}_{\mathcal{A}_\Phi}$. This explains why cocycle attractors are defined as they are, i.e., by using pull-back convergence. For constant stepsize sequences, Theorem 9.1 reduces to results in [43], the starting point of the theory on numerical attractors.

The dual concept to cocycle attractors is that of a cocycle repeller. Reversing time, Theorem 9.1 establishes the existence of a cocycle repeller $R(\mathbf{h})$. Nevertheless, even

for discretizations of Kolmogorov type, the general theory says only that $R(\mathbf{h}) \to Y$ in an upper semicontinuous way. However, in the case that permanence is granted by the standard GALF assumption, $R(\mathbf{h}) = Y$ for every doubly infinite stepsize sequence with $\|\mathbf{h}\|$ sufficiently small.

Actually, a stronger result—a discretized version of Theorem 3.1—holds true.

THEOREM 9.3. *Assume that $P(x) = \prod_{i=1}^n x_i^{p_i}$ is a GALF and let φ be a $\mathcal{P}$th order discretization operator of Kolmogorov type for (1) on X. Then there exist an open neighborhood $\mathcal{W}$ of Y in X and positive constants $h_0, \lambda_1, \lambda_2, \lambda_3$ with the properties as follows. Given an infinite stepsize sequence $\{h_k\}_{k=1}^\infty$ with $\sum_{k=1}^\infty h_k = \infty$ and $\sup h_k \le h_0$,*

$$d_E\left(\{\varphi^{-1}(h_1,\cdot)\circ\cdots\circ\varphi^{-1}(h_M,\cdot)\}(x), Y\right) \le \lambda_1 e^{-\lambda_2(h_1+\cdots+h_M)}(d_E(x,Y))^{\lambda_3}$$

whenever $x \in \mathcal{W}$, $M = 1, 2, \ldots$. Here of course $\varphi^{-1}(h_k,\cdot)$ denotes the inverse of $\varphi(h_k,\cdot)$, $k = 1, 2, \ldots, M$, established by Lemma 7.2.

Proof. The proof is an expanded version of that of Theorem 3.1. For details, see [21]. □

A similar result holds true for asymptotically autonomous systems. Consider the ordinary differential equation

$$\dot{x}_i = x_i e_i(t,x), \qquad (t,x) \in \mathbb{R}\times X, \tag{49}$$

where $e_i : \mathbb{R}\times X \to X$ is a continuous function satisfying

$$e_i(t,x) \to f_i(x) \qquad \text{uniformly in } \ x \in X \ \text{ as } \ t \to \infty, \quad i = 1, 2, \ldots, n$$

and $\sum_i x_i e_i(t,x) = 0$ for each $(t,x) \in \mathbb{R}\times X$. Assume that system (49) has the uniqueness property and that function f in the limiting autonomous system (1) is Lipschitz. The solution of (49) through $(t_0,x) \in \mathbb{R}\times X$ is denoted by $\Psi(\cdot, t_0, x)$.

If the limiting autonomous system (1) is robustly permanent due to a standard GALF, then (49) is permanent too. More precisely, the following result holds true.

THEOREM 9.4. *Assume that $P(x) = \prod_{i=1}^n x_i^{p_i}$ is a GALF for (1) on X. Let $\mathcal{U}_{\mathcal{A}_\Phi}$ be an open neighborhood of $\mathcal{A}_\Phi$ in X and let C be a compact subset of $X \setminus Y$. Given an initial time t_0 arbitrarily, there exists a time T such that*

$$\Psi(t_0+t, t_0, x) \in \mathcal{U}_{\mathcal{A}_\Phi} \qquad \text{whenever } \ t \ge T \ \text{ and } \ x \in C.$$

Proof. The proof is a simple variation of the proof of Theorem 9.3. Some details are contained in [21]. □

10. Connections to index theories. Let K be a nonempty Φ-invariant compact subset of Y. Following Szymczak, Wojcik, and Zgliczynski [63], we say that K is *of repelling type* if $\{x \in X : \emptyset \ne \omega(x) \subset K\} \subset Y$. In view of Theorem 5.2 above, the existence of a GALF for (1) on K implies that K is of repelling type and, for some $\eta > 0$, $\mathcal{B}(K,\eta) \setminus Y$ does not contain entire trajectories. Starting from property $(\beta)_K$, the very same conclusions are derived in the first part of the proof of Theorem 4.4 of Schreiber [57]. By $(\alpha)_K \Leftrightarrow (\beta)_K$ in Theorem 5.3, property $(\beta)_K$ means that $P_K : X \to \mathbb{R}$, $P_K = \prod_{i=1}^n x_i^{p_i}$ defines a GALF for Φ on K. In particular, the existence of a local GALF plus the isolatedness of K with respect to the boundary flow $\Phi|_Y$ imply that K is isolated (i.e., isolated with respect to the entire flow Φ on X).

From now on, assume that $\emptyset \ne K \subset Y$ is a compact isolated invariant set of repelling type. Assume, in addition, that K is a repeller for $\Phi|_Y$. In view of the

Zubov–Ura–Kimura theorem, K is a repeller for Φ (i.e., an attractor for the backward flow Φ^* defined by $\Phi^*(t,x) = \Phi(-t,x)$ for all $(t,x) \in \mathbb{R} \times X$) and thus $\emptyset \neq \alpha(x) \subset K$ for any $x \in X$ with $d_E(x,K)$ sufficiently small. Alternatively, assume that K is an attractor for $\Phi|_Y$. Applying the Zubov–Ura–Kimura theorem again, we find *there exists an* $x \in X \setminus Y$ *with* $\emptyset \neq \alpha(x) \subset K$. Geometrically, the property italicized above means that K repels a trajectory from Y into $X \setminus Y$. However, if K is neither a repeller nor an attractor for $\Phi|_Y$, then the existence of an $x \in X \setminus Y$ with $\emptyset \neq \alpha(x) \subset K$ is a rather delicate matter and requires methods of algebraic topology.

By using standard degree theory, the same problem in $\mathbb{R}^{n-1} \times [0,\infty)$ (and $K \subset \mathbb{R}^{n-1} \times \{0\}$ being a finite collection of equilibria (and $\mathbb{R}^{n-1} \times \{0\}$ invariant)) was first investigated by Hofbauer [29]. Capietto and Garay [13] used the fixed point index (which is an appropriate version of degree theory) and a more general index theory developed by Conley [12]. Their approach, however, worked only for flows induced by vector fields and some special kinds of isolated invariant sets. Both restrictions were removed and much more Conley-type results proved by Wojcik [65]. Generalizations for discrete-time semidynamical systems were given by Szymczak, Wojcik, and Zgliczynski [63]. For details, in particular for the index theories involved, we refer to the original papers [13], [65], [63] and the references cited therein.

The next theorem is a straightforward consequence of the main results of [63] within the context of the present paper.

THEOREM 10.1. *Let $\emptyset \neq K \subset Y$ be a compact isolated invariant set of repelling type. Assume that the homotopical Conley index $I_C(K, \Phi|_Y, Y)$ of K with respect to the boundary flow $\Phi|_Y$ in Y is nontrivial. Then there exists an $x \in X \setminus Y$ with $\emptyset \neq \alpha(x) \subset K$.*

Proof. If $K = Y$, then the Zubov–Ura–Kimura theorem applies.

If $K \neq Y$, then consider a point $y_0 \in Y \setminus K$ and note that the pair $(X \setminus \{y_0\}, Y \setminus \{y_0\})$ is homeomorphic to the pair $(\mathbb{R}^{n-2} \times [0,\infty), \mathbb{R}^{n-2} \times \{0\})$. Modifying the dynamics in a small vicinity of y_0 in X, we may assume that y_0 is an equilibrium point for Φ. Hence all results in [63] (proved for compact isolated invariant subsets of $\mathbb{R}^{n-2} \times \{0\}$, the boundary of the half-space $(\mathbb{R}^{n-2} \times [0,\infty))$ translate into results on compact Φ-invariant subsets of $Y \setminus \{y_0\}$.

By Theorem 2 of [63], the homotopical Conley index $I_C(K, \Phi(1,\cdot), X)$ of K with respect to the time-one map of Φ in X is trivial. If $x \in X \setminus Y$ with $\emptyset \neq \alpha(x) \subset K$ for no $x \in X \setminus Y$, then K is also of attracting type, and thus, by Theorem 1 of [63], $I_C(K, \Phi(1,\cdot), X) = I_C(K, \Phi(1,\cdot)|_Y, Y)$. Since the index map is homotopic to the identity, we conclude that, together with $I_C(K, \Phi(1,\cdot)|_Y, Y)$, also $I_C(K, \Phi|_Y, Y)$ is trivial, a contradiction. □

Unfortunately, it is in general very difficult to check whether the homotopical Conley index $I_C(K, \Phi|_Y, Y)$ is nontrivial or not. Note, however, that nontriviality of $I_C(K, \Phi|_Y, Y)$ is a consequence of $I_F(K, \Phi|_Y, Y) \neq 0$, nontriviality of the fixed point index, and that this latter condition can be fairly easily checked [1]. Besides, $I_F(K, \Phi^*|_Y, Y) = (-1)^n I_F(K, \Phi|_Y, Y)$. The time-duality problem for the homotopical Conley index, in particular the question of whether nontriviality of $I_C(K, \Phi^*|_Y, Y)$ is equivalent to the nontriviality of $I_C(K, \Phi|_Y, Y)$, seems to be open. The answer is affirmative on the homology–cohomology level of the Conley index [53].

The following result is a discretization analogue of Theorem 10.1.

THEOREM 10.2. *Let $\emptyset \neq K \subset Y$ be a compact isolated invariant set of repelling type. Let U be an open neighborhood of K in $\mathbb{R}^n_+$ and let $P_K : U \to \mathbb{R}$, $x \to \prod_{i=1}^n x_i^{p_i}$ be a GALF for* (1) *on K. In addition, assume that K is the maximal compact $\Phi|_Y$-invariant set in* $\mathrm{cl}(U) \cap Y$ *and that the cohomological Conley index $i_C(K, \Phi|_Y, Y)$ is*

nontrivial. Finally, let $K_h \subset Y$ denote the maximal compact $\varphi(h,\cdot)|_Y$-invariant set in $\mathrm{cl}(U) \cap Y$. Then, for h sufficiently small, $K_h \neq \emptyset$ and, for some $x_h \in X \setminus Y$ suitably chosen, $\emptyset \neq \alpha_{\varphi(h,\cdot)}(x_h) \subset K_h$.

Proof. If $K = Y$ and h is small enough, then $K_h = Y$ is a repeller for $\varphi(h,\cdot)$ by Theorem 8.3, and the discrete-time version of the Zubov–Ura–Kimura theorem applies.

If $K \neq Y$ and h is small enough, then $i_C(K_h, \varphi(h,\cdot)|_Y, Y) = i_C(K, \Phi|_Y, Y)$ by the main result in Mrozek and Rybakowski [54] (when applied to $Y \setminus \{y_0\}$, which is locally Lipschitz homeomorphic to $\mathbb{R}^{n-2}$ for any $y_0 \in Y \setminus K$). Hence $K_h \neq \emptyset$. On the other hand, $K_h \subset U \cap Y$ by the upper semicontinuity result in [22], h sufficiently small. Furthermore, combining the proofs of Theorems 5.2 and 8.3, it is not hard to show that K_h (as a subset of X) is isolated with respect to $\varphi(h,\cdot)$ and, for each $x \in U \setminus Y$, inclusion $\emptyset \neq \omega_{\varphi(h,\cdot)}(x) \subset K_h$ is impossible. Thus $\emptyset \neq K_h \subset Y$ is a compact isolated $\varphi(h,\cdot)$-invariant set of repelling type and (as a weakening of the nontriviality of the cohomological Conley index $i_C(K_h, \varphi(h,\cdot)|_Y, Y)$), the homotopical Conley index $I_C(K_h, \varphi(h,\cdot)|_Y, Y)$ is nontrivial. Using Theorems 1 and 2 of [63], the desired result follows immediately. □

Discretizations have better topological properties than general discrete-time dynamical systems. For example, they preserve orientation. Near transversal sections, discretizations (for h small enough) embed to continuous-time local dynamical systems [17], [23]. A further nontrivial topological property of discretizations is what we called numerical Wazewski property [22], [18]. In certain applications, as it was pointed out by Conley [12] himself, the classical Wazewski principle is stronger than the index.

Conjecture. *We conjecture that the nontriviality of the cohomological Conley index $i_C(K, \Phi|_Y, Y)$ in Theorem* 10.2 *can be replaced by the following requirement: Assume that B^+ is not a retract of B, where $B \subset U$ is an isolating block with respect to the boundary flow $\Phi|_Y$ for K in Y and B^+ denotes the entry set of B.*

Combining Theorems 1 in [13] and C4 in [22], we find that the conjecture holds true under the additional conditions that K is contained in a single face of Y and B is an isolating block with corners. One of the major difficulties in proving the conjecture is constructing C^∞ Liapunov functions for attractor–repeller pairs on manifolds with corners (such as Y).

11. Applications. In all the previous sections we worked with the simplex as the phase space and noted only that analogous results are valid in $\mathbb{R}^n_+$ for dissipative flows. In the present section we give applications to systems on other phase spaces such as compact smooth manifolds, half-spaces, products of a simplex with a ray, and products of simplices.

Subsection 11.1 is devoted to differential equations near compact smooth codimension 1 submanifolds of $\mathbb{R}^n$. Note that a much deeper codimension $k \geq 1$ analysis, based on Oseledec's theory, is given in [3], [11] for diffeomorphisms. It is an open problem to extend the concept of GALF (which is at present a codimension 1 object) to codimension k problems.

In subsections 11.2 and 11.3 we study robust permanence of replicator and Lotka–Volterra equations. The problem of successful invasion is discussed in subsection 11.4. Subsection 11.5 is devoted to discretized game dynamics with variable stepsizes.

11.1. Manifolds with smooth boundary. Let Ω be a bounded open set in $\mathbb{R}^n$ and assume that $Z = \partial\Omega$ is a compact smooth codimension 1 submanifold of $\mathbb{R}^n$. Let U be an open neighborhood of Z in $\mathbb{R}^n$ and let $F : U \to \mathbb{R}^n$ be a C^1 function. Assume

that Z is invariant with respect to the (local) flow Θ of the ordinary differential equation $\dot{x} = F(x)$, $x \in U$. The set of Θ-invariant Borel probability measures on Z is denoted by $\mathcal{M}_\Theta(Z)$. Finally, for $z \in Z$, let $\nu(z) \in \mathbb{R}^n$ denote the outer normal unit vector of Z at z. It is well known that, for some $\varepsilon > 0$, mapping $(-\varepsilon, \varepsilon) \times Z \to U$, $(\lambda, z) \to x = z + \lambda\nu(z)$ is a coordinate transformation. In this new system of normal and tangential coordinates, $\dot{x} = F(x)$, $x \in U$ can be rewritten as the system $\dot{\lambda} = N(\lambda, z)$, $\dot{z} = T(\lambda, z)$, $(\lambda, z) \in (-\varepsilon, \varepsilon) \times Z$. Since $N(0, z) = 0$ for each $z \in Z$, there exists a continuous function $S : (-\varepsilon, \varepsilon) \times Z \to \mathbb{R}$ satisfying $N(\lambda, z) = \lambda S(\lambda, z)$. With $P(x) = \lambda$ (or, equivalently, $P(x) = \lambda^p$ for any $p > 0$) as GALF and applying (the corresponding analogue, with (X, Y) replaced by $(\mathrm{cl}(\Omega), Z)$, of) Theorem 2.2 and Lemmas 4.1 and 4.2, we obtain the following theorem.

THEOREM 11.1. *If the normal Liapunov exponent*

$$\int_Z S(0, z)\, d\mu > 0 \tag{50}$$

for each (ergodic) $\mu \in \mathcal{M}_\Theta(Z)$, then Z is a repeller for Θ.

It is not hard to compute $S(0, z)$ explicitly:

$$S(0, z) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{dF_i}{dx_j}(z) \cdot \nu_j(z) \cdot \nu_i(z) = \langle \nu(z), \mathcal{J}(z)\nu(z) \rangle,$$

where $\mathcal{J}(z)$ denotes the Jacobian of F evaluated at z and $\langle \cdot, \cdot \rangle$ denotes the standard scalar product in $\mathbb{R}^n$.

We note that an analogous result holds for abstract manifolds X with smooth collared boundary $\partial X = Z$. Condition (50) implies via Theorem 3.4 the existence of a Liapunov function in a neighborhood of Z of the form $V(x) = \lambda Q(\lambda, z)$ with Q positive and C^1.

Using Pesin's theory, one can show, as in [3] or [57], that the converse of Theorem 11.1 is "almost" true: If F is C^2 and the reverse inequality holds in (50) for at least one invariant measure $\mu \in \mathcal{M}_\Theta(Z)$ (i.e., at least one normal Liapunov exponent is negative), then the invariant manifold Z attracts at least one orbit from $\Omega \setminus Z$. If F is only C^1, then a weaker converse result can be obtained from the ergodic closing lemma, as in [30], [57]: There are arbitrarily small C^1-perturbations of the flow, with Z as invariant manifold, that have a periodic orbit in Z, which has negative normal Floquet exponent and is therefore normally attracting. Other converse results can be derived from index theory, as used in section 10.

We finally remark that Theorem 11.1 can be applied also to study dissipativity in a suitable compactification of the state space and to investigate critical cases of stability by analyzing homogeneous differential equations that arise as the principal part of normal forms, such as Molchanov's theorem [42]; for details see [21].

11.2. Consequences for replicator equations. The replicator equation (4) on the simplex X enjoys an important averaging principle (see [35, Thm. 7.6.4], [41]), which can be stated in terms of time averages or space averages.

LEMMA 11.2. (1) *Suppose that for $x \in X$, $\omega(x)$ is contained in some (relatively open) face of the simplex. Then every limit point of the time average of this solution,*

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T \Phi(t, x) dt, \tag{51}$$

is an equilibrium point $\bar{x}$ on this face.

(2) *Let $\mu \in \mathcal{M}_{\mathcal{F}}^{E}$ be an ergodic invariant measure for* (4). *Then its mean $\bar{x} = \int x d\mu(x)$ is an equilibrium of* (4), *and the external Liapunov exponents of μ coincide with that at $\bar{x}$.*

A consequence of this averaging property is that in applying Theorem 4.4 for replicator equations one can restrict oneself to (convex combinations of) point measures instead of all invariant measures. Hence a finitely computable sufficient condition for robust permanence can be established. In particular, Corollary 2.4 simplifies to the following result from [35] and [41], but now strengthened with robustness.

THEOREM 11.3. *If there are $p_i > 0$ $(i = 1, \dots, n)$ such that for every fixed point $\bar{x}$ of* (4) *in Y,*

$$p \cdot A\bar{x} > \bar{x} \cdot A\bar{x},$$

then $P(x) = \prod_i x_i^{p_i}$ is a GALF, and hence (4) *is robustly permanent.*

Consider now the discrete-time replicator dynamics

$$(\mathcal{F}(x))_i = x_i \frac{1 + h(Ax)_i}{1 + h\ x \cdot Ax}. \tag{52}$$

Here $h > 0$ is such that $1 + ha_{ij} > 0$ for all i, j. Then the map (52) is a diffeomorphism on X [49]. Another discrete-time replicator dynamics is

$$(\mathcal{F}(x))_i = x_i \frac{e^{h(Ax)_i}}{\sum_j x_j\ e^{h(Ax)_j}}, \tag{53}$$

which is a diffeomorphism for h small and, in contrast to (52), enjoys a similar averaging property to that of (4) [35, p. 79, Ex. 7.6.6]. Note that both (52) and (53) are first order discretization operators of Kolmogorov type for (4), but none of them is of the form of those investigated in Example 7.4 (except for zero-sum games, i.e., $A = -A^T$).

The first part of our next result is a simple consequence of Theorem 8.3, whereas the stronger result in the second part follows from Theorem 6.3 and the averaging property. Finding a finitely computable condition for robust permanence for (52) for arbitrary $h > 0$ is an open problem.

THEOREM 11.4. *Let the assumption of Theorem* 11.3 *hold, i.e., there exists a standard ALF $P(x) = \prod_i x_i^{p_i}$ for the continuous-time replicator dynamics* (4). *Then for small $h > 0$, the discrete-time replicator equation* (52) *is robustly permanent. Similarly, for all $h > 0$, P is also an ALF for* (53), *and hence* (53) *is robustly permanent.*

11.3. Consequences for Lotka–Volterra equations. Lotka–Volterra systems

$$\dot{x}_i = x_i\big(r_i + (Ax)_i\big), \quad x \in \mathbb{R}^n_+, \tag{54}$$

enjoy a similar averaging property to that of Lemma 11.2 for replicator equations. This goes essentially back to Volterra; see [35], [41], [57, Lem. 7.1]. Hence for Lotka–Volterra equations (54) it is sufficient to consider point measures at boundary equilibria when applying our permanence results. In particular, our Theorem 4.4 shows that the sufficient condition in Schreiber's [57, Thm. 7.2] for robust permanence of Lotka–Volterra equations is *equivalent* to the sufficient condition for permanence based on a standard GALF due to Jansen [41] (see also [35, Ex. 13.6.3]):

There are $p_i > 0$ $(i = 1, \dots, n)$ such that for every fixed point $\bar{x}$ of (54) *on* $\partial\mathbb{R}^n_+$,

$$p\cdot(r + A\bar{x}) > 0. \tag{55}$$

Now we turn to discrete-time Lotka–Volterra systems such as a higher-dimensional version of the logistic map,

$$(\mathcal{F}(x))_i = x_i(1 + h(r_i + (Ax)_i)), \tag{56}$$

and an exponential version (see [33] and [50]),

$$(\mathcal{F}(x))_i = x_i e^{h(r_i+(Ax)_i)}. \tag{57}$$

The first part of our next result is a simple consequence of Theorem 8.3, whereas the second part follows from Theorem 6.3 and from the averaging principle in [33].

THEOREM 11.5. *Suppose* (56) *and* (57) *are (robustly) dissipative and* (55) *holds true; i.e., there exists a standard ALF $P(x) = \prod_i x_i^{p_i}$ for* (54). *Then for small $h > 0$,* (56) *is (robustly) permanent. Similarly, for all $h > 0$, P is an ALF for* (57), *which is (robustly) permanent.*

Dissipativity of (57) is discussed in [33] and [50].

As a further application we rederive and strengthen a recent result of Mierczyński and Schreiber [52] on totally permanent Lotka–Volterra systems. They established the equivalence (L2) $\Leftrightarrow$ (L4) with their weaker meaning of robust permanence. Note that (L3) and (L4) are computable conditions.

THEOREM 11.6. *The following conditions are equivalent:*

(L1) *Equation* (54) *as well as all its subsystems are permanent.*

(L2) *Equation* (54) *as well as all its subsystems are robustly permanent.*

(L3) *$-A$ is a P-matrix (i.e., all principal minors of $-A$ are positive) and each (relatively open k-dimensional, $k = 1, 2, \dots, n$) face of $\mathbb{R}^n_+$ contains an equilibrium.*

(L4) *Equation* (54) *is dissipative, each face contains a unique equilibrium, and all its external eigenvalues are positive.*

Proof. Every permanent Lotka–Volterra system has a unique interior equilibrium and $\det(-A) > 0$; see [35, Thms. 13.5.1 and 13.5.2]. Applying this to all subsystems we conclude that (L1) implies (L3). The P-matrix property implies the dissipativity of (54) and also uniqueness of saturated equilibrium; see [35, Thms. 15.2.1 and 15.4.5]. Hence no boundary equilibrium can have an external eigenvalue ≤ 0. This shows (L3) $\Rightarrow$ (L4). Finally (L4) implies (L2) by Corollary 4.6, and (L2) $\Rightarrow$ (L1) is trivial. ∎

11.4. Invasion of a permanent system. Consider a permanent n-species community and a further species which is able to invade that resident community. Will the invader be able to survive, i.e., will the population move towards a new stable community consisting of the invader and a certain subset of the resident population? In the biological literature, e.g., in [66], this question is often phrased as, *Does invasion lead to persistence?*

A positive answer to this question is possible only under stringent assumptions. For example, if in the resident system there are several attractors, and the new species invades at one attractor, it could be driven out again by leading the population to the other attractor. This can be avoided only if all normal Liapunov exponents on the global interior attractor are positive.

Even then, one could imagine that the population evolves to a state where the invader as well as some of the resident species are eliminated. A simple example in two dimensions is the system

$$\dot{x} = x\left(x(1-x) - y\right),$$
$$\dot{y} = y(x-y).$$

The density of the invading species y increases near the resident equilibrium $x = 1, y = 0$. But every interior solution converges to the origin: $\frac{y}{x}$ increases monotonically, and for $y > x$, y decreases. However, this dynamics is degenerate, since the origin is not hyperbolic. If the resident system is robustly permanent (thanks to a GALF), this extinction phenomenon cannot occur. This is the essence of the next theorem, which generalizes an analogous result on Lotka–Volterra equations in [32].

THEOREM 11.7. *We consider a system of n resident species on $X = \mathbb{R}^n_+$ and an invader whose density we denote by $y \geq 0$,*

$$\dot{x}_i = x_i f_i(x, y), \tag{58}$$
$$\dot{y} = y g(x, y) \tag{59}$$

on the augmented state space $X' = \mathbb{R}^n_+ \times \mathbb{R}_+$. We identify X with the subsystem $X \times \{0\}$ of X' and assume that (58)–(59) give rise to a dissipative dynamical system $\Theta = (\Phi, \Psi) : \mathbb{R} \times X' \to X' = X \times \mathbb{R}_+$. In addition, assume there exists a GALF for the resident system (58) with $y = 0$ which is therefore robustly permanent. Finally, assume that the global attractor $\mathcal{A} \subset \operatorname{int} X$ of (58) is nonsaturated in the sense that

$$\int_{\mathcal{A}} g(x,0)\, d\mu_x > 0 \qquad \text{for each} \quad \mu_x \in \mathcal{M}_\Theta(\mathcal{A}). \tag{60}$$

Then

$$\limsup_{t\to\infty} \Psi(t, x, y) > 0 \qquad \text{for all } (x, y) \in \operatorname{int} X'.$$

Proof. To the contrary, suppose that $\Psi(t, z, w) \to 0$ for some $(z, w) \in \operatorname{int} X'$. Hence $\emptyset \neq \omega_\Theta(z, w) \subset X$. Actually, since $(\mathcal{A}, \partial X)$ is an attractor–repeller decomposition of the resident system, $\omega_\Theta(z, w) \subset \mathcal{A}$ or $\omega_\Theta(z, w) \subset \partial X = Y$. Combining Theorems 5.3 and 5.2, condition (60) implies that the first inclusion is impossible. Hence $\omega_\Theta(z, w) \subset Y$.

Consider now the GALF $P : X \to \mathbb{R}$ and observe for each $t \in \mathbb{R}$ that

$$\frac{d}{dt}\log(P(\Phi(t, z, w))) = \sum_{i=1}^{n} p_i(\Phi(t, z, w)) f_i(\Phi(t, z, w), \Psi(t, z, w)).$$

Integrating between 0 and T yields

$$\frac{1}{T}\log\left(\frac{P(\Phi(T, z, w))}{P(z)}\right) = \frac{1}{T}\int_0^T \sum_{i=1}^{n} p_i(\Phi(t, z, w)) f_i(\Theta(t, z, w))\, dt. \tag{61}$$

Since $\Phi(t, z, w)$ approaches Y, $P(\Phi(T, z, w)) \to 0$ as $T \to \infty$. Thus each limit point of the left-hand side of (61) is ≤ 0. On the other hand, property $\Psi(T, z, w) \to 0$, the GALF assumption, Lemma 4.2, and the robustness arguments in the proof of

Theorem 2.2 imply that every limit point as $T \to \infty$ on the right-hand side is positive. This is a contradiction. □

An obvious drawback of Theorem 11.7 is that only positivity of the limsup can be guaranteed. However, one cannot do better in general: It is easy to construct examples with $\liminf_{t\to\infty} \Psi(t,x,y) = 0$. Consider (58)–(59) on $X' = \mathbb{R}^4_+$ with three resident species (1,2,3). Suppose, that the invader eliminates resident species 3 and forms together with 1 and 2 a system with an attracting heteroclinic cycle, such as in Example 2.6. Then $\limsup_{t\to\infty} \Psi(t,x,y) > 0$ but $\liminf_{t\to\infty} \Psi(t,x,y) = 0$. Moreover, the invader gets arbitrarily close to 0 for arbitrarily long times. Hence, practically, the invader is not safe from extinction. This example shows that information on the global dynamics of the full system (58)–(59) is needed to guarantee persistence after invasion.

11.5. Discretizations and diminishing stepsizes. Hofbauer and Schlag [34] studied imitation dynamics for two-person (bimatrix) games. These led to recurrence relations of the following form ($k = 0, 1, \dots$):

$$\begin{aligned} p_i^{k+1} &= p_i^k \left(1 + h f_i\left(p^k, q^k\right)\right), \\ q_j^{k+1} &= q_j^k \left(1 + h g_j\left(p^k, q^k\right)\right). \end{aligned} \tag{62}$$

Here the state space X is a product of two simplices, $X = X_n \times X_m \subseteq \mathbb{R}^n_+ \times \mathbb{R}^m_+$, and $f_i, g_j : X_n \times X_m \to \mathbb{R}$ ($1 \le i \le n$, $1 \le j \le m$) are appropriately chosen functions such that $(p^k, q^k) \mapsto (p^{k+1}, q^{k+1})$ defines a map from $X_n \times X_n$ into itself. In the limit $h \to 0$ these discrete-time models tend to the differential equation

$$\begin{aligned} \dot{p}_i &= p_i f_i(p,q), \\ \dot{q}_j &= q_j g_j(p,q). \end{aligned} \tag{63}$$

The functions f_i, g_j are given by

$$\begin{aligned} f_i(p,q) &= (\pi^1(i,q) - \pi^1(p,q))\phi^1(\pi^1(p,q)), \\ g_j(p,q) &= (\pi^2(p,j) - \pi^2(p,q))\phi^2(\pi^2(p,q)), \end{aligned} \tag{64}$$

where π^1, π^2 are the payoff functions for the two players, and ϕ^i are strictly decreasing functions with positive values.

For $n = m = 2$, i.e., each of the two players has two pure strategies, X is simply the square $[0,1]^2$. Of particular interest are games with a cyclic structure: For these games, the boundary of the square, Y, forms a heteroclinic cycle for the dynamics (62) and (63). They have a unique Nash equilibrium E which lies in the interior of X and which has been shown [34, Thm. 1] to be globally asymptotically stable for (63); i.e., E is the dual attractor to the repeller Y. Furthermore, there exists a standard GALF for (63), so that Y is a robust repeller.

For small enough $h \in (0, h_0)$ (h_0 being the minimal slope of the reciprocal of ϕ^i), the map (62) still has a standard GALF, so that Y remains a robust repeller; see [34, Prop. 1]. The dual attractor, $\mathcal{A}_h$, contains E (which is unstable for the maps [34, Prop. 3]) as proper subset.

Combining now Theorem 8.3 and the upper semicontinuity result [20] for the attractor–repeller pair (E, Y) of (63) shows that for small enough $h > 0$, the numerical attractor $\mathcal{A}_h$ arising from E attracts all of the interior of the square X, i.e., is dual to the robust repeller Y. This confirms the conjecture in [34, p. 535, footnote 4].

On the other hand, if instead of a constant stepsize h a decreasing sequence of stepsizes satisfying $h_k \to 0$ and $\sum h_k = \infty$ is chosen in (62), then every orbit starting

in the interior of the square converges to the equilibrium E. This conjecture from [34, p. 539 and footnote 9] follows now from the permanence result in Theorem 9.3 above, Lemma 4 from [20] (or Benaim and Hirsch [6]), and the attractor–repellor decomposition (E, Y) for the limiting differential equation (63).

REFERENCES

[1] O. Aberth, *Computation of topological degree using interval arithmetic and applications*, Math. Comp., 62 (1994), pp. 171–178.

[2] E. Akin, *The General Topology of Dynamical Systems*, AMS, Providence, RI, 1978.

[3] P. Ashwin, J. Buescu, and I. N. Stewart, *From attractor to chaotic saddle: A tale of transverse instability*, Nonlinearity, 9 (1996), pp. 703–737.

[4] A. Babin and M. I. Vishik, *Attractors of Evolution Equations*, North–Holland, Amsterdam, 1992.

[5] R. W. Bass, *Zubov's stability criterion*, Bol. Soc. Mat. Mexicana, 4 (1959), pp. 26–29.

[6] M. Benaim and M. W. Hirsch, *Asymptotic pseudotrajectories and chain recurrent flows, with applications*, J. Dynam. Differential Equations, 8 (1996), pp. 141–176.

[7] W. J. Beyn and J. Lorenz, *Center manifolds of dynamical systems under discretization*, Numer. Funct. Anal. Optim., 9 (1987), pp. 381–414.

[8] N. P. Bhatia and G. P. Szegö, *Stability Theory of Dynamical Systems*, Springer-Verlag, Berlin, 1970.

[9] A. Bressan, *High order approximation of implicitly defined maps*, Ann. Mat. Pura Appl., 137 (1984), pp. 163–173.

[10] F. E. Browder, *A further generalization of the Schauder fixed point theorem*, Duke Math. J., 32 (1965), pp. 575–578.

[11] J. Buescu, *Exotic Attractors: From Liapunov Stability to Riddled Basins*, Birkhäuser, Boston, 1997.

[12] C. Conley, *Isolated Invariant Sets and the Morse Index*, AMS, Providence, RI, 1978.

[13] A. Capietto and B. M. Garay, *Saturated invariant sets and the boundary behaviour of differential systems*, J. Math. Anal. Appl., 176 (1993), pp. 166–181.

[14] M. Dellnitz and O. Junge, *On the approximation of complicated dynamical behaviour*, SIAM J. Numer. Anal., 36 (1999), pp. 491–515.

[15] H. I. Freedman and P. Waltman, *Persistence in models of three interacting predator–prey populations*, Math. Biosci., 68 (1984), pp. 213–231.

[16] B. M. Garay, *Uniform persistence and chain recurrence*, J. Math. Anal. Appl., 139 (1989), pp. 372–381.

[17] B. M. Garay, *The discretized flow on domains of attraction: A structural stability result*, IMA J. Numer. Anal., 18 (1998), pp. 77–90.

[18] B. M. Garay, *Some remarks on Wazewski's retract principle*, Univ. Iagel. Acta Math., 36 (1998), pp. 97–105.

[19] B. M. Garay, *A functional equation characterizing monomial functions used in permanence theory for ecological differential equations*, Univ. Iagel. Acta. Math., submitted.

[20] B. M. Garay and J. Hofbauer, *Chain recurrence and discretization*, Bull. Austral. Math. Soc., 55 (1997), pp. 63–71.

[21] B. M. Garay and J. Hofbauer, *Robust Permanence for Ecological Differential Equations, Minimax, and Discretizations*, preprint, 2001; available online from http://mailbox.univie.ac.at/Josef.Hofbauer/galf.htm.

[22] B. M. Garay and P. E. Kloeden, *Discretization near compact invariant sets*, Random Comput. Dynam., 5 (1997), pp. 93–123.

[23] B. M. Garay and P. L. Simon, *Numerical flow–box theorems under structural assumptions*, IMA J. Numer. Anal., 21 (2001), pp. 733–749.

[24] A. Gaunersdorfer, *Time averages for heteroclinic attractors*, SIAM J. Appl. Math., 52 (1992), pp. 1476–1489.

[25] L. Grüne, *Convergence rates of perturbed attracting sets under vanishing perturbation*, J. Math. Anal. Appl., 244 (2000), pp. 369–392.

[26] M. W. Hirsch, H. L. Smith, and X.-Q. Zhao, *Chain transitivity, attractivity and strong repellors for semidynamical systems*, J. Dynam. Differential Equations, 13 (2001), pp. 107–131.

[27] J. Hofbauer, *A general cooperation theorem for hypercycles*, Monatsh. Math., 91 (1981), pp. 233–240.

[28] J. Hofbauer, *A unified approach to persistence*, Acta Appl. Math., 14 (1989), pp. 11–22.
[29] J. Hofbauer, *An index theorem for dissipative semiflows*, Rocky Mountain J. Math., 20 (1990), pp. 1017–1031.
[30] J. Hofbauer, *Permanence on a Halfspace and the Ergodic Theorem*, manuscript, presented at SIAM meeting, Orlando, FL, 1990.
[31] J. Hofbauer, *Heteroclinic cycles in ecological differential equations*, Tatra Mt. Math. Publ., 4 (1994), pp. 105–116.
[32] J. Hofbauer, *Invasion, Permanence and Heteroclinic Cycles*, preprint, 1998.
[33] J. Hofbauer, V. Hutson, and W. Jansen, *Coexistence for systems governed by difference equations of Lotka-Volterra type*, J. Math. Biol., 25 (1987), pp. 553–570.
[34] J. Hofbauer and K. Schlag, *Sophisticated imitation in cyclic games*, J. Evol. Economics, 10 (2000), pp. 523–543.
[35] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.
[36] J. Hofbauer and J. W.-H. So, *Uniform persistence and repellors for maps*, Proc. Amer. Math. Soc., 107 (1989), pp. 1137–1142.
[37] V. Hutson, *A theorem on average Liapunov functions*, Monatsh. Math., 98 (1984), pp. 267–275.
[38] V. Hutson, *The stability under perturbations of repulsive sets*, J. Differential Equations, 76 (1988), pp. 77–90.
[39] V. Hutson and K. Mischaikow, *An approach to practical persistence*, J. Math. Biol., 37 (1998), pp. 447–466.
[40] V. Hutson and W. Moran, *Persistence of species obeying difference equations*, J. Math. Biol., 15 (1982), pp. 203–213.
[41] W. Jansen, *A permanence theorem for replicator and Lotka Volterra systems*, J. Math. Biol., 25 (1987), pp. 411–422.
[42] L. G. Khazin and E. E. Shnol, *Stability of Critical Equilibrium States*, Manchester University Press, Manchester, UK, 1991.
[43] P. E. Kloeden and J. Lorenz, *Stable attracting sets in dynamical systems and their one-step discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 986–995.
[44] P. E. Kloeden and B. Schmalfuss, *Lyapunov functions and attractors under variable time–step discretization*, Discrete Contin. Dynam. Systems, 2 (1996), pp. 163–172.
[45] P. E. Kloeden and B. Schmalfuss, *Nonautonomous systems, cocycle attractors and variable time-step discretization*, Numer. Algorithms, 14 (1997), pp. 141–154.
[46] N. Krylov and N. Bogoliubov, *Les mesures invariantes et transitives dans la mécanique non linéaire*, Mat. Sb. (N.S.), 1 (1936), pp. 707–710.
[47] M. K. Kwong and A. Zettl, *Norm inequalities for derivatives and differences*, in Inequalities, W. N. Everett, ed., Marcel Dekker, Basel, 1991, pp. 91–121.
[48] M. C. Li, *Structural stability under numerics*, J. Differential Equations, 141 (1997), pp. 1–12.
[49] V. Losert and E. Akin, *Dynamics of games and genes: Discrete versus continuous time*, J. Math. Biol., 17 (1983), pp. 241–251.
[50] Z. Lu and W. Wang, *Permanence and global attractivity for Lotka-Volterra difference systems*, J. Math. Biol., 39 (1999), pp. 269–282.
[51] R. Mañé, *Ergodic Theory and Differentiable Dynamics*, Springer-Verlag, Berlin, 1987.
[52] J. Mierczyński and S. J. Schreiber, *Kolmogorov vector fields with robustly permanent subsystems*, J. Math. Anal. Appl., 267 (2002), pp. 329–337.
[53] M. Mrozek and R. Srzednicki, *On time-duality of the Conley index*, Results Math., 24 (1993), pp. 161–167.
[54] M. Mrozek and K. P. Rybakowski, *Discretized ordinary differential equations and the Conley index*, J. Dynam. Differential Equations, 4 (1992), pp. 57–63.
[55] V. V. Nemytzkii and V. V. Stepanov, *Qualitative Theory of Differential Equations*, Princeton University Press, Princeton, NJ, 1960.
[56] S. Schreiber, *On growth rates of subadditive functions for semiflows*, J. Differential Equations, 148 (1998), pp. 334–350.
[57] S. Schreiber, *Criteria for C^r robust permanence*, J. Differential Equations, 162 (2000), pp. 400–426.
[58] P. Schuster, K. Sigmund, and R. Wolff, *Dynamical systems under constant organization. III. Cooperative and competitive behaviour of hypercycles*, J. Differential Equations, 32 (1979), pp. 357–368.
[59] S. Simmons, *Minimax and Monotonicity*, Springer-Verlag, Berlin, 1998.
[60] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK, 1996.

[61] G. P. Szegö and G. Trecanni, *Semigruppi di Transformationi Multivoche*, Springer-Verlag, Berlin, 1969.
[62] W. Szlenk, *An Introduction to the Theory of Smooth Dynamical Systems*, John Wiley, New York, 1984.
[63] A. Szymczak, K. Wojcik, and P. Zgliczynski, *On the discrete Conley index in the invariant subspace*, Topology Appl., 87 (1998), pp. 105–115.
[64] T. Ura and I. Kimura, *Sur le courant extérieur à une région invariante. Théorème de Bendixson*, Comm. Math. Univ. Sanctii Pauli, 8 (1960), pp. 23–39.
[65] K. Wojcik, *An attraction result and an index theorem for continuous flows on* $\mathbb{R}^n \times [0, \infty)$, Ann. Polon. Math., 65 (1997), pp. 203–211.
[66] G. Wolkowicz, *Invasion of a persistent system*, Rocky Mountains J. Math., 20 (1990), pp. 1217–1234.
[67] T. Yoshizawa, *Stability Theory by Lyapunov's Second Method*, Math. Soc. Japan, Tokyo, 1966.
[68] V. I. Zubov, *The Methods of A. M. Liapunov and Their Application*, Izdatelstvo Leningradskogo Universiteta, Moskva, 1957 (in Russian).
[69] V. I. Zubov, *The Methods of A. M. Liapunov and Their Application*, Noordhoff, Groningen, 1964.

SIAM J. MATH. ANAL.
Vol. 34, No. 5, pp. 1040–1063

FINITE-DIMENSIONAL ATTRACTORS AND EXPONENTIAL ATTRACTORS FOR THE NAVIER–STOKES EQUATIONS OF COMPRESSIBLE FLOW*

DAVID HOFF† AND MOHAMMED ZIANE‡

Abstract. We prove that the uniform attractor for the Navier–Stokes equations of compressible flow with quasi-periodic external forces has finite fractal dimension. As a byproduct of our analysis, we also obtain the existence of finite-dimensional *exponential* attractors for the Navier–Stokes system.

Key words. Navier–Stokes equations, compressible flow, exponential attractors

AMS subject classifications. 35Q30, 35B41, 76N10

PII. S0036141002400889

1. Introduction. We prove the finite dimensionality of the compact attractor obtained in [11], as well as the existence of finite-dimensional exponential attractors, for the Navier–Stokes equations of compressible fluid flow in one space dimension,

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2)_x + P(\rho)_x = \varepsilon u_{xx} + \rho f, \qquad 0 < x < L, \end{cases} \tag{1.1}$$

with boundary and initial conditions

$$\begin{cases} u(0,t) = u(L,t) = 0, \\ (\rho, u)(\cdot, 0) = (\rho_0, u_0), \\ \displaystyle\int_0^L \rho_0 \, dx = M. \end{cases} \tag{1.2}$$

Here $\rho(x,t)$ and $u(x,t)$ are the density and fluid velocity, $P(\rho) = c^2\rho$ is the isothermal pressure, ε is a viscosity constant, and $f = f(x,t)$ is a time-dependent external force. The attractor, which is determined by the force f, is the smallest closed set in (ρ, u) space which is invariant for the flow and which attracts all trajectories as time goes to infinity. The exponential attractor is a compact, positively invariant subset of (ρ, u) space which contains the attractor, which attracts all trajectories exponentially in time, and which has finite fractal dimension. Precise definitions of the uniform and exponential attractors are given in Definitions 3.1 and 3.2, and the fractal dimension is defined, for example, in [17].

In [10] and [11] we gave a complete global existence, uniqueness, continuous dependence, and regularity theory for solutions of (1.1)–(1.2), obtaining *time-independent* estimates for solutions with large, discontinuous initial data and large external forces. We found, however, that, due to the persistence of singularities in solutions, continuous dependence on initial data would not hold in any reasonable topology which dominates the sup norm of ρ. We therefore constructed the attractor in a reduced

*Received by the editors January 12, 2002; accepted for publication (in revised form) September 6, 2002; published electronically April 15, 2003.
http://www.siam.org/journals/sima/34-5/40088.html
†Department of Mathematics, Indiana University, Bloomington, IN 47405 (hoff@indiana.edu).
The research of this author was supported in part by NSF grant DMS-9986658.
‡Department of Mathematics, University of Southern California, Los Angeles, CA 90089 (ziane@math.usc.edu).

phase space $Y \times X_u$, where $X_u = L^2$ and Y is the set of positive BV densities ρ for which the Lebesgue decomposition of the distribution ρ_x has a singular part consisting of a countable sum of point masses and an absolutely continuous part in L^2. We let $\mathcal{U}_f(t,\tau)$ be the semiprocess on the phase space $Y \times X_u$ defined by taking $\mathcal{U}_f(t,\tau)(\rho_0,u_0)$ to be the solution of (1.1)–(1.2) with force f and initial condition (ρ_0,u_0) imposed at time τ. Then assuming that f belongs to a compact subset Σ of $W^{1,\infty}([0,\infty);\, L^2) \cap L^\infty([0,\infty); W^{1,\infty})$ and that there exists a continuous family of translation operators $\mathcal{T} :\ \Sigma \mapsto \Sigma$, we proved the continuity and asymptotic compactness properties required for the existence of a uniform compact attractor $\mathcal{A}_\Sigma \subset Y \times X_u$, and we showed that $\mathcal{A}_\Sigma$ is a compact subset of $H^1 \times H^1_0$ and is contained in $H^1 \times H^2 \cap H^1_0$. Finally, we proved that the attractor is characterized by a finite number of so-called determining nodes, but the question of finite dimensionality of the attractor was left open.

The purpose of the present paper is therefore to establish that, for quasi-periodic external forces, the uniform attractor $\mathcal{A}_\Sigma$ obtained in [11] does in fact have finite fractal dimension. As a byproduct of our analysis, we will also obtain the existence of finite-dimensional *exponential* attractors for the Navier–Stokes system.

Our specific assumptions on the external force f are as follows: Let $\mathcal{C}^1_b(\mathbb{R}; W^{1,\infty}([0,1]))$ be the set of bounded functions with bounded first derivatives, and assume that

$$
\begin{aligned}
&f \in \mathcal{C}^1_b(\mathbb{R}; W^{1,\infty}([0,1])), \quad f(\cdot,t) = \tilde{f}(\cdot,\alpha_1 t,\dots,\alpha_N t),\\
&\text{where } \alpha_1,\dots,\alpha_N \text{ are rationally independent, and} \qquad (1.3)\\
&\tilde{f}(\cdot,\omega_1,\dots,\omega_N) \text{ is } 2\pi\text{-periodic in each argument } \omega_i,\ i = 1,\dots,N.
\end{aligned}
$$

We fix $\tilde{f}$ and denote by Σ the set of forces f satisfying the above conditions. This set Σ will be fixed throughout the paper and is a particular realization of the more general classes Σ for which global attractors were constructed in [11].

We shall show that the uniform attractor $\mathcal{A}_\Sigma$ has finite fractal dimension with respect to $H^r \times H^r_0$ for a particular $r \in (0,1)$. This means that the balls which are counted in coverings of $\mathcal{A}_\Sigma$ are chosen in the topology of $H^r \times H^r_0$. We also recall a classical result of Mañe [13], who proves that any metric space with fractal dimension less than $m_0/2$ can be embedded in $\mathbb{R}^{m_0}$ with a Lipschitz function whose inverse is Hölder continuous (see [8]). Consequently, the uniform attractor $\mathcal{A}_\Sigma$ for the Navier–Stokes system (1.1)–(1.2) can be embedded in $\mathbb{R}^{m_0}$ for some m_0 depending on the size of the force and the constants appearing in the equations.

The following theorem contains the main results of this paper.

THEOREM 1.1. *The Navier–Stokes system* (1.1)–(1.2) *with external force f satisfying* (1.3) *and with data in the phase space $Y \times X_u$ (of discontinuous functions) has a uniform attractor $\mathcal{A}_\Sigma$ which is compact in $H^1 \times H^1_0$ and which has finite fractal dimension with respect to the $H^r \times H^r_0$-topology for some $r \in (0,1)$. The fractal dimension is bounded above by a constant depending only on N, ε, L, M, and the size of the force. In addition, there exists a compact set $\mathcal{M}_\Sigma$ in $H^1 \times H^1_0$ which is positively invariant under the family of semiprocesses $\mathcal{U}_f(t,\tau)$, contains the attractor $\mathcal{A}_\Sigma$, has finite fractal dimension with respect to the $H^r \times H^r_0$-topology, and has the property that trajectories of solutions starting at initial points in $(Y \times X_u) \cap (H^1 \times H^1_0)$ converge exponentially as time goes to infinity to $\mathcal{M}_\Sigma$ in the topology of $H^r \times H^r_0$.*

We now give a brief outline of the ideas in the proof of Theorem 1.1. First, following [3], we imbed the Navier–Stokes equations (1.1)–(1.2) in a family of equations depending on a parameter σ in the N-dimensional torus $\mathbb{T}^N$, now taking $f(\cdot,t) =$

$\tilde f(\cdot,\omega(t))$, where $\omega(t) = (\alpha t + \sigma)(\mathrm{mod}\ \mathbb{T}^N)$ and $\alpha = (\alpha_1,\dots,\alpha_N)$. The set of such forces is thus parameterized by σ, and the associated processes will now be denoted by $\mathcal{U}_\sigma(t,\tau)$. That is, $\mathcal{U}_\sigma(t,\tau) = \mathcal{U}_f(t,\tau)$ for $f(\cdot,t) = \tilde f(\cdot,\alpha t + \sigma(\mathrm{mod}\mathbb{T}^N))$. To the family $\mathcal{U}_\sigma(t,\tau)$ of semiprocesses we associate the operator $S(t)$ defined by

$$\begin{aligned} &S(t) : (Y\times X_u)\times\mathbb{T}^N \to (Y\times X_u)\times\mathbb{T}^N,\\ &\qquad (\rho,u;\sigma)\mapsto S(t)(\rho,u;\sigma) = (\mathcal{U}_\sigma(t,0)(\rho,u);(\alpha t+\sigma)(\mathrm{mod}\ \mathbb{T}^N)). \end{aligned} \tag{1.4}$$

The set of operators $S(t)$ thus forms a semigroup on $Y\times X_u\times\mathbb{T}^N$. In [11] we showed that this semigroup has a *global* attractor $\mathcal{A}$ (see Definition 3.4 below) which is compact in $H^1\times H^1_0\times\mathbb{T}^N$, and hence compact in $Y\times X_u\times\mathbb{T}^N$. The *uniform* attractor $\mathcal{A}_\Sigma$ referred to in Theorem 1.1 associated with the family $\mathcal{U}_\sigma(t,\tau)$ is then obtained by projecting the global attractor $\mathcal{A}$ onto $Y\times X_u$. Thus if $\mathcal{A}$ has finite fractal dimension, then so does $\mathcal{A}_\Sigma$.

One standard approach to finite dimensionality is through the Kaplan–Yorke trace formula, which requires time-independent bounds for solutions of the system obtained by linearizing about a trajectory in the attractor. As we pointed out in [11], however, such bounds appear to be unavailable for these particular equations. We shall instead prove that $\mathcal{A}_\Sigma$ and $\mathcal{A}$ have finite fractal dimension by applying the following result of Ladyzhenskaya [12].

LEMMA 1.2. *Let M be a subset of a Hilbert space H and assume that M is invariant under a map $S : M\mapsto H$, that is, that $M = S(M)$. Assume also that there is a positive integer n and constants $C>0$ and $\delta\in(0,1)$ such that, for any two points w_1, w_2 in the set M,*

$$||S(w_1)-S(w_2)||_H \le C||w_1-w_2||_H, \tag{1.5}$$

$$||Q_nS(w_1)-Q_nS(w_2)||_H \le \delta||w_1-w_2||_H, \tag{1.6}$$

where Q_n is the projector onto a subspace of H of codimension n. Then the set M has finite fractal dimension bounded by a constant $d(C,\delta)n$, which is proportional to n.

We shall apply this result with $H = H^r\times H^r\times\mathbb{T}^N$ for a particular $r\in(0,1)$, $M=\mathcal{A}$, and $S = S(t^*)$, where t^* is large. To verify the hypotheses (1.5) and (1.6) we apply a decomposition method introduced in [2]. Specifically, we shall show that, for $w_1, w_2\in H$, there is a decomposition $S(t)w_1 - S(t)w_2 = W^I(t) + W^{II}(t)$, with W^I satisfying

$$|W^I(t)|_{H^r\times H^r_0\times\Sigma} \le K_0e^{-\nu_1 t}|w_1-w_2|_{H^r\times H^r_0\times\Sigma} \tag{1.7}$$

for positive constants ν_1 and K_0 depending only on the size of the attractor, and W^{II} satisfying

$$|W^{II}(t)|_{H^1\times H^1_0\times\Sigma} \le K_0e^{\nu_2 t}|w_1-w_2|_{H^r\times H^r_0\times\Sigma}. \tag{1.8}$$

We let V_n be the vector space generated by $\{e^{2k\pi ix/L}\}_{|k|\le n}$ and take Q_n to be the projector onto the orthogonal complement of $V_n\times V_n\times\mathbb{T}^n$ in $H^1\times H^1_0\times\mathbb{T}^N$. Then by (1.8),

$$|Q_nW^{II}(t)|_{H^r\times H^r_0\times\Sigma} \le \frac{C}{n^{1-r}}|W^{II}(t)|_{H^1\times H^1_0\times\Sigma} \le \frac{CK_0}{n^{1-r}}e^{\nu_2 t}|w_1-w_2|_{H^r\times H^r_0\times\Sigma} \tag{1.9}$$

for a constant C. The coefficients on the right-hand sides of (1.7) and (1.9) can then be made arbitrarily small by choosing $t = t^*$ sufficiently large, then n sufficiently large depending on t^*. This proves (1.6), and (1.5) is easy to check; the finite dimensionality of $\mathcal{A}$, and hence of $\mathcal{A}_\Sigma$, then follows.

This same decomposition can also be applied to show that the semigroup $S(t)$ satisfies the so-called squeezing property, and that this squeezing property implies the existence of exponential attractors, as described in the statement of Theorem 1.1. Since the exponential attractor has finite fractal dimension and contains the uniform attractor, its existence affords a second (though not unrelated) proof that the uniform attractor has finite fractal dimension.

The main contribution of the present paper is therefore the construction of the decomposition $S(t)w_1 - S(t)w_2 = W^I(t) + W^{II}(t)$ and the derivation of its essential properties. This decomposition is rather complicated, even for the relatively simple system (1.1)–(1.2) (see the description in section 3.2), and the analysis leading to the required bounds (1.7) and (1.8) is based on a fairly involved sequence of energy estimates for L^2 and H^1 norms, together with a Riesz–Thorin interpolation argument for the intermediate H^r norms.

This paper is organized as follows. First, in section 2 we recall the results of [11] concerning the existence, uniqueness, and continuous dependence of solutions of the system (1.1)–(1.2) (but only for relatively smooth initial data, which is all that is required here), as well as results establishing the existence of the attractor $\mathcal{A}_\Sigma$. We also derive several higher-order regularity estimates which were not given in [11] but which will be needed here. In section 3.1 we describe the abstract framework in which the proof of Theorem 1.1 will be given, including precise definitions of uniform and exponential attractors and statements of known results concerning the squeezing property and the related decomposition of the semigroup. In section 3.2 we construct this decomposition for the particular semigroup associated with the solution operator of (1.1)–(1.2) and then apply the abstract results of section 3.1 to complete the proof of Theorem 1.1. The essential properties of this decomposition are stated and applied in section 3.2, but the derivations, which are rather long and technical, are deferred to section 4.

Existence and regularity of solutions of the Navier–Stokes equations (1.1)–(1.2) in one space dimension have been studied by a great many authors. See the references in [11], for example, for a representative list. The unique feature of the existence theory of [11] is that solutions are constructed satisfying *time-independent* estimates, even with large external forces and large, discontinuous initial data. This, together with the uniqueness and continuous dependence results of [11], makes an attractor theory both possible and interesting. A corresponding well-posedness theory for the case of several space variables is not yet available, however, so that an attractor theory, at least in the sense of the present paper, is not yet feasible. On the other hand, there is a weaker notion of attractor, whose existence does not depend on uniqueness and continuous dependence, and which has been shown to exist for the Navier–Stokes equations in three space dimensions; see Feireisl [7].

The notion of exponential attractor was introduced in [4], and the overall approach of the present paper has also been applied successfully in a number of other contexts. See, for example, [6], [9], [14], and [15].

2. Preliminaries. In this section we recall the results of [10] and [11] concerning the existence and regularity properties of solutions of (1.1)–(1.2) and the existence of the global and uniform attractors. We also derive several additional higher-order regularity estimates required for the analysis in sections 3 and 4.

First we write the Navier–Stokes equations of compressible fluid flow in one space dimension in nondimensional form:

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2)_x + A^2 \rho_x = u_{xx} + \rho f \ , \quad 0 < x < 1, \\ u(0,t) = u(1,t) = 0, \\ (\rho, u)(\cdot, 0) = (\rho_0, u_0), \\ \displaystyle\int_0^1 \rho_0 \ dx = 1, \end{cases} \tag{2.1}$$

where A is the dimensionless constant $A = cM/\varepsilon$.

We say that (ρ, u) is a weak solution of (2.1) on $[0,T]$ provided that

$$(\rho, u) \ \in \ C([0,T];\ L^2([0,1])) \text{ with } (\rho, u)(\cdot, 0) = (\rho_0, u_0), \tag{2.2}$$

$$u \ \in \ C((0,T];\ H_0^1([0,1])), \tag{2.3}$$

$$\rho \ \in \ L^\infty([0,1] \times [0,T]) \quad \text{and} \quad \rho > 0 \text{ a.e.;} \tag{2.4}$$

for test functions $\varphi \ \in \ C^1([0,1] \times [0,T])$ and for times $t_1, t_2 \ \in [0,T]$,

$$\int_0^1 (\rho\varphi)(x,\cdot) dx \ \Big|_{t_1}^{t_2} - \int_{t_1}^{t_2} \int_0^1 (\rho\varphi_t + \rho u \varphi_x) dx dt = 0; \tag{2.5}$$

and for $\psi \ \in \ C^1([0,1] \times [0,T])$ with $\psi(0,t) = \psi(1,t) = 0$,

$$\begin{aligned} &\int_0^1 (\rho u \psi)(x,\cdot) dx \ \Big|_{t_1}^{t_2} - \int_{t_1}^{t_2} \int_0^1 (\rho u \psi_t + \rho u^2 \psi_x + A^2 \rho \psi_x) dx dt \\ &\qquad = \int_{t_1}^{t_2} \int_0^1 (-u_x \psi_x + \rho f \psi) dx dt \,. \end{aligned} \tag{2.6}$$

We define

$$G(\rho) = \rho \log \rho - \rho + 1, \tag{2.7}$$

and for weak solutions (ρ, u) of (2.1) with ρ in H^1,

$$H(\rho, u) = \int_0^1 \left[\tfrac{1}{6} \ \rho u^2 + A^2 G(\rho) + \tfrac{1}{16} \ \frac{\rho_x^2}{\rho^3} \right] dx. \tag{2.8}$$

($H(\rho, u)$ is thus a function of t.) We shall denote the usual norm in $L^2([0,1])$ by $|\cdot|$.

The following theorem gives a version of the existence, uniqueness, and regularity theory of [10] and [11] for the special case that the initial density is in H^1.

THEOREM 2.1. (a) *Given positive constants C_0 and C_Σ, arbitrarily large, there is a constant $C = C(C_0, C_\Sigma, A)$ such that if $f \in W^{1,\infty}([0,\infty); L^2)$ with*

$$\sup_{t \geq 0} [|f(\cdot,t)| + |f_t(\cdot,t)|] \ \leq C_\Sigma, \tag{2.9}$$

and if (ρ_0, u_0) satisfies $\rho_0 > 0$ a.e., $\rho_0, \rho_0^{-1} \in L^\infty$, $\int_0^1 \rho_0 \ dx = 1$, and

$$H(\rho_0, u_0) \leq C_0, \tag{2.10}$$

then there is a unique global weak solution of (2.1) *satisfying* (2.2)–(2.6)*, as well as*

$$H(\rho(\cdot,t),u(\cdot,t)) \le C, \quad t \ge 0, \tag{2.11}$$

$$|u(\cdot,t)| + (1\wedge t)^{1/2}|u_x(\cdot,t)| + (1\wedge t)|u_t(\cdot,t)| \le C, \quad t>0, \tag{2.12}$$

(*where* $1\wedge t \equiv \min\{1,t\}$)*, and*

$$\int_t^{t+1}\int_0^1 [u_x^2 + (1\wedge s)u_t^2 + (1\wedge s)^2 u_{xt}^2]dxds \ \le C, \quad t>0. \tag{2.13}$$

(b) *There is a constant* K_f*, depending only on* C_Σ *and on* A*, such that, given constants* C_0 *and* τ*, there is a time* $T = T(C_0, C_\Sigma, \tau, A)$*, so that if* (ρ,u) *is a weak solution of* (2.1) *with force* f*, satisfying* (2.9)–(2.13) *with constants* C_0, C_Σ*, then for* $t \ge T$*,*

$$H(\rho(\cdot,t),u(\cdot,t)) \le K_f. \tag{2.14}$$

(c) *The solution transforms to Lagrangian coordinates: that is, if we define the Lagrangian coordinate*

$$h(x,t) = \int_0^x \rho(s,t)ds,$$

its inverse $\Phi(h,t)$*, given by*

$$\Phi(h(x,t),t) = x,$$

and functions

$$\begin{aligned} v(h,t) &= \rho(\Phi(h,t),t)^{-1},\\ w(h,t) &= u(\Phi(h,t),t), \end{aligned} \tag{2.15}$$

then (v,w) *is a weak solution of the corresponding system*

$$\begin{cases} v_t - w_h = 0\\ w_t + (A^2 v^{-1})_h = \left(\dfrac{w_h}{v}\right)_h + f\circ\Phi\ , \quad 0<h<1,\\ w(0,t) = w(1,t) = 0. \end{cases} \tag{2.16}$$

For the sake of simplicity, in this paper we will use the Lagrangian formulation and assume throughout that u_τ and v_τ are given and satisfy

$$u_\tau \in L^2, \quad v_\tau \in H^1 \quad \text{with} \quad v_\tau > 0,\ v_\tau \in L^\infty, \tag{2.17}$$

and that the force f satisfies $f \in C_b^1(\mathbb{R}^+; L^2)$. Furthermore, we will assume that f is quasi-periodic in time. More precisely

$$f(\cdot,t) = \tilde f(\cdot, \alpha_1 t, \dots, \alpha_N t), \tag{2.18}$$

where $\tilde{f}(\cdot,\omega_1,\dots,\omega_N)$ is 2π-periodic in each argument ω_i and the α_i are rationally independent. We now replace h by x and rewrite the Navier–Stokes equations in Lagrangian coordinates

$$v_t = u_x, \tag{2.19}$$

$$u_t + A^2\big(v^{-1}\big)_x = \Big(\frac{u_x}{v}\Big)_x + f(\Phi(x,t),t), \tag{2.20}$$

with the boundary conditions

$$u(0,t) = u(1,t) = 0, \quad \int_0^1 v(x,t)\,dx = 1 \tag{2.21}$$

and initial data, given at time $t = \tau$,

$$u(x,\tau) = u_\tau, \quad v(x,\tau) = v_\tau. \tag{2.22}$$

We have the following as a consequence of Theorem 2.1.

THEOREM 2.2. *Let C_0 and C_Σ be two positive constants, arbitrarily large. Then there exist positive constants a_0, b_1, α, β depending on C_Σ and A such that if $f \in W^{1,\infty}(\mathbb{R}_+, L^2)$ is given satisfying $\|f\|_{W^{1,\infty}(\mathbb{R}_+,L^2)} \le C_\Sigma$, and if initial data (v_τ, u_τ) is given satisfying $v_\tau > 0$, $\int_0^1 v_\tau(x)dx = 1$, and*

$$|u_\tau| + |(v_\tau)_x| + |{v_\tau}^{-1}|_{L^\infty} \le C_0, \tag{2.23}$$

then there exists a unique global weak solution (v,u) satisfying

$$v \in C([\tau,\infty); H^1), \quad v(\cdot,\tau) = v_\tau, \quad v(\cdot,t) > 0,\ t \ge \tau, \tag{2.24}$$

$$u \in C([\tau,\infty); L^2) \cap C((\tau,\infty); H^2 \cap H_0^1); \quad u(\cdot,\tau) = u_\tau. \tag{2.25}$$

Furthermore, there exists $T_0 = T_0(C_0, C_\Sigma, A)$ such that, for $t - \tau \ge T_0$, we have

$$|u(\cdot,t)|^2 \le a_0^2, \quad \int_t^{t+1} |u_x|^2 ds \le b_1^2, \quad |v_x(\cdot,t)| \le \beta, \quad \frac{1}{\alpha} \le v(x,t) \le \alpha. \tag{2.26}$$

Thus if we define $\mathcal{B} \subset H^1 \times H_0^1$ to be the set of (v,u) satisfying (2.26) *(so that $\mathcal{B}$ is determined solely by C_Σ and A), then given C_0, there is a time T_0 depending only on C_0, C_Σ, and A such that if $(v(\cdot,t),u(\cdot,t))$ is the solution of* (2.19)–(2.22) *with initial data (v_τ,u_τ) satisfying the conditions $v_\tau > 0$, $\int_0^t v_\tau(x)dx = 1$, and* (2.23)*, then $(v(\cdot,t),u(\cdot,t)) \in \mathcal{B}$ for $t \ge T_0(C_0, C_\Sigma, A)$.*

The set $\mathcal{B}$ is called the absorbing ball. It is the first indication of a dissipative mechanism for the compressible Navier–Stokes equations. Its existence, together with the asymptotic compactness property of the solution operator, leads to the existence of the uniform attractor; see [10] and [11] for more details. Theorem 2.2 allows us to define a family of semiprocesses $\{\mathcal{U}_f(t,\tau)\}_{t\ge\tau\ge 0}$ on the set $\mathcal{X} = \{(v,u) \in H^1 \times L^2, v > 0, v^{-1} \in L^\infty\}$ as follows:

$$\begin{aligned} &\mathcal{U}_f(t,\tau)\colon\ \mathcal{X} \to \mathcal{X}, \\ &\qquad (v,u) \mapsto \mathcal{U}_f(t,\tau)(v,u) = (v(\cdot,t), u(\cdot,t)), \end{aligned} \tag{2.27}$$

where $(v(\cdot,t),u(\cdot,t))$ is the solution of (2.19)–(2.22) at time t with initial data (v,u) given at time τ.

We also established in [10, 11], using Theorem 2.2, the existence of the uniform attractor for the family of semiprocesses $\{\mathcal{U}_f(t,\tau)\}$. The precise statement is as follows.

THEOREM 2.3. *Let C_Σ and A be as in Theorem 2.2. Then there is a set $\mathcal{A}$ which is compact in $H^1 \times H_0^1 \times \Sigma$, which is invariant under $S(t)$, and which attracts all trajectories $S(t)w_0$ for $w_0 \in \mathcal{X} \times \Sigma$ in the topology of $H^1 \times H_0^1 \times \Sigma$. The projection of $\mathcal{A}$ onto $H^1 \times H_0^1$ is compact in the topology of $H^1 \times H_0^1$, is positively invariant under the family of semiprocesses $\mathcal{U}_f(t,\tau)$, and is uniformly attracting in the sense that, for all bounded sets $B \subset \mathcal{X}$,*

$$\lim_{t\to\infty} \sup_{f\in\Sigma} \operatorname{dist}_{\mathcal{X}}(\mathcal{U}_f(t,\tau)B, \mathcal{A}_\Sigma) = 0 \qquad \forall\ \tau \geq 0. \tag{2.28}$$

The sets $\mathcal{A}$ and $\mathcal{A}_\Sigma$ are, respectively, the global attractor for the semigroup $S(t)$ and the uniform attractor for the family of semiprocesses $\mathcal{U}$; see section 3 for more formal definitions.

The remainder of this section is devoted to obtaining uniform estimates on some higher derivatives of the solution of (2.19)–(2.22). In deriving these estimates, we will make use of the uniform Gronwall lemma (see [17, page 89]): if $g,\ h$, and y are three positive locally integrable function on (t_0,∞) such that $\frac{dy}{dt} \leq gy + h$ for $t \geq t_0$, and if

$$\int_t^{t+1} g(s)\,ds \leq k_1, \quad \int_t^{t+1} h(s)\,ds \leq k_2, \quad \int_t^{t+1} y(s)\,ds \leq k_3 \quad \text{for } t \geq t_0,$$

where $k_1,\ k_2,\ k_3$ are positive constants, then

$$y(t+1) \leq (k_2 + k_3)\exp(k_1) \quad \text{for } t \geq t_0 + 1.$$

(i) Uniform estimates on $|u_x|$. We multiply (2.20) by u_{xx} and integrate with respect to x to obtain

$$\frac{1}{2}\frac{d}{dt}|u_x|^2 + \int_0^1 \frac{{u_{xx}}^2}{v}\,dx = \int_0^1 \frac{u_x u_{xx} v_x}{v^2}\,dx - A^2\int_0^1 \frac{v_x}{v^2}u_{xx}\,dx - \int_0^1 f u_{xx}\,dx, \tag{2.29}$$

and thanks to (2.26), we obtain

$$\frac{d}{dt}|u_x|^2 + \frac{1}{\alpha}|u_{xx}|^2 \leq C\alpha^{11}\beta^4|u_x|^2 + CA^4\alpha^5\beta^2 + C\alpha C_\Sigma^2, \tag{2.30}$$

where C will denote here and throughout this paper a numerical constant. Therefore, by the uniform Gronwall lemma and (2.13), we obtain

$$|u_x(\cdot,t)|^2 \leq a_1^2 \quad \text{and} \quad \int_t^{t+1} |u_{xx}|^2 ds \leq b_2^2 \quad \text{for } t - \tau \geq T_0 + 1, \tag{2.31}$$

where

$$a_1^2 = \left[b_1^2 + A^4\alpha^5\beta^2 + \alpha C_\Sigma^2\right]\exp(C\alpha^{11}\beta^4), \tag{2.32}$$

$$b_2^2 = C(\alpha^{12}\beta^4 b_1^2 + A^4\alpha^6\beta^2 + \alpha^2 C_\Sigma^2) + 2\alpha a_1^2. \tag{2.33}$$

Furthermore, using (2.20), we have

$$|u_t|^2 \leq 2A^4\alpha^4|v_x|^2 + 2|u_{xx}|^2\alpha^2 + 2\alpha^2|u_x||u_{xx}||v_x|^2 + 2C_\Sigma^2. \tag{2.34}$$

Therefore

$$\int_t^{t+1} |u_t|^2\,ds \le 2(A^4\alpha^4\beta^2 + b_2^2 + \alpha^2 b_2^2 + 2\alpha^4\beta^2 b_1 b_2) + 2C_\Sigma^2 = b_3^2. \tag{2.35}$$

(ii) Estimates on $|u_t|$. Taking the derivative with respect to time in (2.20), multiplying by u_t, and integrating over $(0,1)$, we obtain after a few calculations

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}|u_t|^2 + \int_0^1 \frac{u_{xt}^2}{v^2}\,dx &= -A^2\int_0^1 \frac{u_x u_{xt}}{v^2}\,dx - \int_0^1 \frac{u_x^2 u_{xt}}{v^2}\,dx + \int_0^1 \frac{df}{dt}u_t\,dx \\ &\le A^2\alpha^2 a_1|u_{xt}| + \alpha^2|u_x||u_{xx}||u_{xt}| + \left|\frac{df}{dt}\right||u_{xt}|.\end{aligned}$$

Hence

$$\frac{d}{dt}|u_t|^2 + \frac{1}{\alpha}|u_{xt}|^2 \le CA^4\alpha^5 a_1^2 + 2\alpha^5 a_1^2|u_{xx}|^2 + 2\alpha C_\Sigma^2, \tag{2.36}$$

and since $u_{xx} = vu_t - A^2\frac{v_x}{v} + \frac{u_x v_x}{v} - vf$, we have

$$|u_{xx}|^2 \le 2\alpha^2|u_t|^2 + 2A^4\alpha^2\beta^2 + 2\alpha^2\beta^2|u_x||u_{xx}| + 2\alpha^2 C_\Sigma^2, \tag{2.37}$$

and, using Young's inequality together with (2.31), we obtain

$$|u_{xx}|^2 \le 16\alpha^2|u_t|^2 + 16\alpha^2(A^4\alpha^2\beta^2 + \alpha^2\beta^4 a_1^2 + C_\Sigma^2). \tag{2.38}$$

Hence

$$\frac{d}{dt}|u_t|^2 + \frac{1}{\alpha}|u_{tx}|^2 \le k_1^2|u_t|^2 + k_2^2, \tag{2.39}$$

where

$$k_1^2 = C\alpha^7 a_1^2, \tag{2.40}$$

$$k_2^2 = C\alpha^7 a_1^2(A^4\beta^2 + \alpha^2\beta^4 a_1^2 + C_\Sigma^2) + 2A^4\alpha^5 a_1^2 + 2\alpha C_\Sigma^2. \tag{2.41}$$

Applying the uniform Gronwall lemma, we obtain

$$|u_t(\cdot,t)|^2 \le (k_2^2 + b_3^2)\exp(k_1^2) = a_2 \quad \text{for } t - \tau \ge T_0 + 1, \tag{2.42}$$

$$\int_t^{t+1} |u_{xt}|^2\,ds \le k_1^2 b_3^2 k_2^2 = b_4^2 \quad \text{for } t - \tau \ge T_0 + 1, \tag{2.43}$$

$$|u_{xx}(\cdot,t)|^2 \le C(k_2^2 + b_3^2)\exp(k_1^2) + 16\alpha^2(A^4\beta^2 + \alpha^2\beta^4 a_1^2 + C_\Sigma^2) = a_2^2. \tag{2.44}$$

(iii) Uniform estimates on $|u_{xt}|$. We differentiate (2.20) with respect to time and multiply by u_{tt} to obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_0^1 \frac{u_{xt}^2}{v}\,dx + |u_{tt}|^2 &= -A^2\int_0^1 \left(\frac{1}{v}\right)_{xt} u_{tt}\,dx - \frac{1}{2}\int_0^1 \frac{u_{xt}^2 u_x}{v^2}\,dx \\ &\quad + \int_0^1 \left(\frac{u_x^2}{v^2}\right)_x u_{tt}\,dx + \int_0^1 u_{tt}\frac{df}{dt}\,dx.\end{aligned} \tag{2.45}$$

Now note that

$$\left(\frac{1}{v}\right)_{xt} = -\frac{u_{xx}}{v^2} + 2\frac{u_x v_x}{v^3}.$$

Hence

$$\left|\left(\frac{1}{v}\right)_{xt}\right|^2 \le 2\alpha^4 a_2^2 + 4\alpha^6 |u_x||u_{xx}||v_x| \le 2\alpha^4 a_2^2 + 4\alpha^6 a_1 a_2 \beta,$$

$$\left|\int_0^1 \frac{u_{xt}^2 u_x}{v^2}\,dx\right| \le \alpha^2 |u_x|_{L^\infty} |u_{xt}|^2 \le \alpha^2 (a_1 a_2)^{1/2} |u_{xt}|^2,$$

and

$$\begin{aligned}\left|\int_0^1 \left(\frac{u_x^2}{v^2}\right)_x u_{tt}\,dx\right| &= \left|\int_0^1 \left[\frac{2u_x u_{xx}}{v^2} u_{tt} - \frac{2u_x^2 v_x}{v^3} u_{tt}\right] dx\right| \\ &\le 2\alpha^2 |u_x|_{L^\infty} a_2 |u_{tt}| + 2\alpha^3 |u_x|_{L^\infty}^2 |u_{tt}|.\end{aligned}$$

Therefore

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_0^1 \frac{u_{xt}^2}{v}\,dx + \frac{1}{2}|u_{tt}|^2 &\le C(\alpha^4 a_2^2 + \alpha^6 a_1 a_2 \beta) A^4 + C(a_1 a_2)^{1/2} |u_{xt}|^2 \\ &\quad + C\alpha^4 a_2^3 a_1 + C\alpha^6 a_1^2 a_2^2 + C C_\Sigma^2 \\ &\le k_3^2 + C\alpha^3 (a_1 a_2)^{1/2} \int_0^1 \frac{u_{xt}^2}{v}\,dx.\end{aligned} \tag{2.46}$$

Hence

$$\frac{d}{dt}\int_0^1 \frac{{u_{xt}}^2}{v}\,dx + |u_{tt}|^2 \le k_3^2 + k_4^2 \int_0^1 \frac{u_{xt}^2}{v}\,dx,$$

where

$$\begin{aligned}k_3^2 &= C(\alpha^4 a_2^2 + \alpha^6 a_1 a_2 \beta) A^4 + C\alpha^4 a_2^3 a_1 + C\alpha^6 a_1^2 a_2^2 + C C_\Sigma^2, \\ k_4^2 &= C\alpha^3 (a_1 a_2)^{1/2}.\end{aligned}$$

Finally, using the uniform Gronwall lemma, we obtain

$$|u_{xt}|^2 \le \alpha (k_3^2 + b_4^2) \exp(k_4^2). \tag{2.47}$$

We collect the previous estimates in the following lemma.

LEMMA 2.4. *Let C_0 and C_Σ be positive constants, arbitrarily large, and let (v, u) be the solution described in Theorem 2.2. Then there exists a positive constant M_0 depending only on C_Σ and A, and there exists a time $T_0 = T_0(C_0, C_\Sigma, A)$ such that if $||f||_{W^{1,\infty}} \le C_\Sigma$ and $t - \tau \ge T_0 + 1$, then*

$$|u(\cdot,t)|^2 + |u_x(\cdot,t)|^2 + |u_{xx}(\cdot,t)|^2 + |u_t(\cdot,t)|^2 + |u_{xt}(\cdot,t)|^2 \le M_0^2. \tag{2.48}$$

In particular, this inequality holds for any solution on the uniform attractor $\mathcal{A}_\Sigma$.

We conclude this section with the following lemma, which is a consequence of Lemma 2.4.

LEMMA 2.5. *Let the hypotheses and notation of Lemma 2.4 be in force, and let g be a smooth function. Then there exists a constant $M_0 = M_0(C_\Sigma, A, g)$, independent of C_0, such that*

$$\int_0^1 \left|\frac{\partial^2}{\partial x \partial t} g(v(x,t))\right|^2 dx + \int_0^1 \left|\frac{\partial^2}{\partial t^2} g(v(x,t))\right|^2 dx \le M_0^2 \tag{2.49}$$

for $t-\tau \geq T_0+1$. In particular, if $(v(\cdot,\cdot),u(\cdot,\cdot))$ and $(\bar{v}(\cdot,\cdot),\bar{u}(\cdot,\cdot))$ are two solutions in $\mathcal{A}_\Sigma$ and if G is a smooth real-valued function, then

$$\theta(x,t) \equiv \frac{G(v)-G(\bar{v})}{v-\bar{v}} = \int_0^1 G'(\tau v(x,t)+(1-\tau)\bar{v}(x,t))\,d\tau \tag{2.50}$$

satisfies

$$|\theta_{xt}(\cdot,t)|^2 + |\theta_{tt}(\cdot,t)|^2 \leq M_0^2 \tag{2.51}$$

for all t.

Proof. We write

$$(g(v))_{xt} = (g'(v)u_x)_x = g''(v)u_x^2 + g'(x)u_{xx}. \tag{2.52}$$

Since, by Lemma 2.4, $|u_{xx}(\cdot,t)|_{L^2}$, $|g''(v(\cdot,t))|_{L^\infty}$, and $|u_x^2(\cdot,t)|_{L^2}$ are bounded for $t-\tau \geq T_0+1$, we have that $|g(v)_{xt}(\cdot,t)|$ is bounded for $t-\tau \geq T$. Similarly

$$(g(v))_{tt} = (g'(v)u_x)_t = g''(v)u_x^2 + g'(v)u_{xt}, \tag{2.53}$$

and, thanks to Lemma 2.4, the proof is complete. ∎

3. Construction of the exponential attractor: Proof of Theorem 1.1. In this section we prove Theorem 1.1 by constructing the decomposition discussed in section 1 for the semigroup associated with the semiprocess generated by the solution operator for the Navier–Stokes system (1.1)–(1.2). This decomposition enables us to apply the result of Lemma 1.2 to show that the attractor $\mathcal{A}$ of Theorem 2.3 has finite fractal dimension, and therefore that its projection $\mathcal{A}_\Sigma$ has finite fractal dimension as well. The same decomposition also enables us to apply other known results to show that the semigroup $S(t)$ satisfies the so-called squeezing property and that, as a consequence, it has an exponential attractor of finite fractal dimension, as does the associated process $\mathcal{U}_\sigma$.

We begin in section 3.1 by giving a brief description of the abstract framework in which the proof of Theorem 1.1 will be given, including the general formulation of the required decomposition of the semigroup and statements of standard results concerning the implications of this decomposition for the existence of exponential attractors. Then in section 3.2 we show how to construct the required decomposition for the semigroup associated with the system (1.1)–(1.2) and how the abstract results of section 3.1 can be applied to complete the proof of Theorem 1.1. The properties of the decomposition required for the application of this abstract framework will be seen to follow from certain a priori bounds for solutions of two systems of differential equations derived from (1.1)–(1.2). These estimates are stated in Theorem 3.9, and their proofs are deferred to section 4.

3.1. Exponential attractors. Let $\mathcal{Y}$ be a metric space and consider the equation

$$\begin{cases} \dfrac{du}{dt} = \mathcal{F}(\alpha t, u), \quad t \geq \tau, \quad \tau \in \mathbb{R}, \\ u(\tau) = u_\tau \in \mathcal{Y}, \end{cases} \tag{3.1}$$

where $\alpha = (\alpha_1,\dots,\alpha_N)$. We assume that the α_i's are rationally independent and that $\mathcal{F}(\omega_1,\dots,\omega_N,\cdot)$ is 2π-periodic in each argument ω_i, $i=1,\dots,N$. In the spirit

of the theory developed in [3] for nonautonomous systems, we associate with (3.1) the following family of equations, depending on a parameter $\sigma \in \mathbb{T}^N$:

$$\begin{cases} \dfrac{du}{dt} = \mathcal{F}(\alpha t + \sigma, u), \quad t \geq \tau, \quad \tau \in \mathbb{R}, \\ u(\tau) = u_\tau \in \mathcal{Y}. \end{cases} \tag{3.2}$$

Assuming that (3.2) is well-posed for initial data in $\mathcal{Y}$, we can define a family of semiprocesses $\mathcal{U}_\sigma(t,\tau)$ by taking $\mathcal{U}_\sigma(t,\tau)u_\tau$ to be the solution of (3.2) at time t. Thus

$$\mathcal{U}_\sigma(t,t) = Id \quad \forall\, t \in \mathbb{R}, \quad \forall\, \sigma \in \mathbb{T}^N, \tag{3.3}$$

$$\mathcal{U}_\sigma(t,s) \circ \mathcal{U}_\sigma(s,\tau) = \mathcal{U}_\sigma(t,\tau) \quad \forall\, t \geq s \geq \tau, \quad \forall\, \sigma \in \mathbb{T}^N. \tag{3.4}$$

Uniform attractors and uniform exponential attractors for semiprocesses are then defined as follows (see [16]).

DEFINITION 3.1. *A closed set $\mathcal{A}_\Sigma \subset \mathcal{Y}$ is called the uniform (with respect to σ) attractor for the family of processes $\mathcal{U}_\sigma(t,\tau)$ defined on $\mathcal{Y}$ if*

(a) *$\mathcal{A}_\Sigma$ is a compact subset of $\mathcal{Y}$;*

(b) *for all $B \subset \mathcal{Y}$ bounded,*

$$\lim_{t\to\infty} \sup_{\sigma\in\mathbb{T}^N} dist_{\mathcal{Y}}(\mathcal{U}_\sigma(t,\tau)B, \mathcal{A}_\Sigma) = 0; \tag{3.5}$$

(c) *for all $\mathcal{A}' \subset \mathcal{Y}$ closed and satisfying* (b), *$\mathcal{A}_\Sigma \subset \mathcal{A}'$.*

DEFINITION 3.2. *Let H and V be Hilbert spaces with V compactly imbedded in H, and assume that the family of semiprocesses $\mathcal{U}_\sigma(t,\tau)$ is defined on a closed subset $\mathcal{Y}$ of V. A closed set $\mathcal{M}_\Sigma \subset \mathcal{Y}$ is a uniform $\mathcal{Y}-V$ exponential attractor for the family of processes $\mathcal{U}_\sigma(t,\tau)$ and for initial data in V if*

(a) *$\mathcal{A}_\Sigma \subset \mathcal{M}_\Sigma \subset V$, where $\mathcal{A}_\Sigma$ is the uniform attractor;*

(b) *$\mathcal{M}_\Sigma$ is compact in V and has finite fractal dimension;*

(c) *for all $B \subset \mathcal{Y}$ bounded in V, there exist constants $c_1(B)$ and $c_2(B)$ such that*

$$\sup_{\sigma\in\mathbb{T}^N} dist_H(\mathcal{U}_\sigma(t,\tau)B, \mathcal{M}_\Sigma) \leq c_1 e^{-c_2(t-\tau)}. \tag{3.6}$$

REMARK 3.3. *Note that the exponential convergence in* (3.6) *is in a weaker norm than that in which the set B is assumed to be bounded. For example, as we shall see below for the Navier–Stokes system* (1.1)–(1.2), *the set $\mathcal{Y}$ will be the closed absorbing ball (see Theorem* 2.3*) in the Hilbert space $V = H^1 \times H^1_0$, whereas H will be the Hilbert space $H^r \times H^r_0$ for some $r \in (0,1)$. Thus exponential convergence to $\mathcal{M}_\Sigma$ will occur in the topology of $H = H^r \times H^r_0$ for sets $B \subset \mathcal{Y}$ which are bounded in the stronger topology of $H^1 \times H^1_0$.*

The proof of the existence of uniform exponential attractors for nonautonomous evolution equations follows the theory in [3]: one considers the semigroup $S(t)$ defined on the extended phase space $\mathcal{Y} \times \mathbb{T}^N$ by

$$\begin{aligned} S(t)\colon\ & \mathcal{Y} \times \mathbb{T}^N \longrightarrow \mathcal{Y} \times \mathbb{T}^N, \\ & (u,\sigma) \longmapsto (\mathcal{U}_\sigma(t,0)u, (\alpha t + \sigma)(\mathrm{mod}\mathbb{T}^N)) \end{aligned} \tag{3.7}$$

and then proves that $S(t)$ has an exponential attractor $\mathcal{M}$. The uniform exponential attractor $\mathcal{M}_\Sigma$ for the semiprocess $\mathcal{U}_\sigma$ is then obtained by projecting $\mathcal{M}$ onto $\mathcal{Y}$. We recall the definitions of global attractor and exponential attractor for semigroups.

DEFINITION 3.4. *A closed set $\mathcal{A} \subset \mathcal{Y} \times \mathbb{T}^N$ is called the global attractor for the semigroup $S(t)$ defined on $\mathcal{Y} \times \mathbb{T}^N$ if*

(a) *$\mathcal{A}$ is a compact subset of $\mathcal{Y} \times \mathbb{T}^N$;*

(b) *$\mathcal{A}$ is invariant under $S(t)$, i.e., $S(t)\mathcal{A} = \mathcal{A}$ for all t;*

(c) *for all $B \subset \mathcal{Y} \times \mathbb{T}^N$ bounded,*

$$\lim_{t\to\infty} dist_{\mathcal{Y}\times\mathbb{T}^N}(S(t)B, \mathcal{A}) = 0. \tag{3.8}$$

DEFINITION 3.5. *Let H and V be Hilbert spaces with V compactly imbedded in H, and assume that $S(t)$ is a semigroup defined on $\mathcal{Y} \times \mathbb{T}^N$, where $\mathcal{Y}$ is a closed subset of V. A closed set $\mathcal{M} \subset V \times \mathbb{T}^N$ is a uniform $\mathcal{Y} \times \mathbb{T}^N - V \times \mathbb{T}^N$ exponential attractor for the semigroup $S(t)$ for initial data in $V \times \mathbb{T}^N$ if*

(a) *$\mathcal{A} \subset \mathcal{M} \subset V \times \mathbb{T}^N$, where $\mathcal{A}$ is the global attractor of $S(t)$;*

(b) *$\mathcal{M}$ is compact in $V \times \mathbb{T}^N$ and has finite fractal dimension;*

(c) *for all $B \subset \mathcal{Y} \times \mathbb{T}^N$ bounded in $V \times \mathbb{T}^N$, there exist constants $c_1(B)$ and $c_2(B)$ such that*

$$\sup dist_{H\times\mathbb{T}^N}(S(t)B, \mathcal{M}) \le c_1 e^{-c_2(t-\tau)}. \tag{3.9}$$

As discussed earlier, an effective method for proving the existence of exponential attractors for semigroups arising from systems which are only partially dissipative, and therefore whose solution operators are only asymptotically compact, is based on a decomposition of the semigroup $S(t)$ into the sum of two operators $S(t) = S_1(t) + S_2(t)$, where $S_1(t)$ is compact in a suitable sense, $S_2(t)$ is continuous, and

$$\sup_{w\in C} |S_2(t)w|_{\mathcal{Y}\times\mathbb{T}^N} \to 0 \quad \text{as} \quad t \to \infty \tag{3.10}$$

for every bounded set $C \subset \mathcal{Y} \times \mathbb{T}^N$. The existence of such a decomposition is known to imply a certain "squeezing property," and this in turn is known to imply the existence of an exponential attractor. This squeezing property, its relation to the above decomposition of the semigroup, and its sufficiency for the existence of exponential attractors are described as follows.

DEFINITION 3.6. *We say that the semigroup $S(t)$ satisfies the squeezing property on a positively invariant closed set $X \subset \mathcal{Y} \times \mathbb{T}^N$ if for every $\delta \in (0, 1/4)$ there exist an orthogonal projector $\mathcal{P}$ with finite rank and a time $t^*(\delta) > 0$ such that for all $(w_1, w_2) \in X \times X$, either*

$$|S(t^*)w_1 - S(t^*)w_2|_{\mathcal{Y}\times\mathbb{T}^N} \le \delta |w_1 - w_2|_{\mathcal{Y}\times\mathbb{T}^N} \tag{3.11}$$

or

$$|(I - \mathcal{P})(S(t^*)w_1 - S(t^*)w_2)|_{\mathcal{Y}\times\mathbb{T}^N} \le |\mathcal{P}(S(t^*)w_1 - S(t^*)w_2)|_{\mathcal{Y}\times\mathbb{T}^N}. \tag{3.12}$$

The following result of Eden et al. [4] guarantees the existence of exponential attractors for semigroups satisfying the squeezing property.

THEOREM 3.7. *If the semigroup $S(t)$ defined by* (3.7) *satisfies the squeezing property of Definition* 3.6*, and if the global attractor $\mathcal{A}$ exists in the sense of Definition* 3.4*, then $S(t)$ has a finite-dimensional exponential attractor $\mathcal{M}$, and the semiprocess $\mathcal{U}_\sigma(t, \tau)$ associated with $S(t)$ has a finite-dimensional exponential attractor $\mathcal{M}_\Sigma$.*

We shall verify that the semigroup associated with the Navier–Stokes system (1.1)–(1.2) does indeed satisfy the squeezing property by applying the following sufficient condition of Galusinski, Hind, and Miranville [9].

PROPOSITION 3.8. *Let V and H be Hilbert spaces with V compactly imbedded in H. Assume that $\mathcal{Y}$ is a closed subset of V and that for each $n \in \mathbb{N}$ there exists an orthogonal projector $P_n\colon\ H \to H$ with finite rank such that*

$$\forall\ y \in V, \quad |Q_n y|_H \le C(n)|y|_V, \tag{3.13}$$

where $Q_n = I - P_n$ and $\lim_{n\to\infty} C(n) = 0$. Assume also that there are continuous functions d and h on $[0,\infty)$ such that $\lim_{t\to\infty} d(t) = 0$, and that for all w_1 and w_2 in $\mathcal{Y} \times \mathbb{T}^N$, there exist $W^I(t)$ and $W^{II}(t)$ such that

$$S(t)w_1 - S(t)w_2 = W^I(t) + W^{II}(t), \ \textit{with} \tag{3.14}$$

$$|W^I(t)|^2_{H\times\mathbb{T}^N} \le d(t)|w_1 - w_2|^2_{H\times\mathbb{T}^N}, \tag{3.15}$$

$$|W^{II}(t)|^2_{V\times\mathbb{T}^N} \le h(t)|w_1 - w_2|^2_{H\times\mathbb{T}^N}. \tag{3.16}$$

Then $S(t)$ satisfies the squeezing property of Definition 3.6.

Proof. See [9]. □

3.2. Application to the Navier–Stokes system: Proof of Theorem 1.1. It suffices to prove Theorem 1.1 for the Lagrangian formulation (2.19)–(2.22) of the Navier–Stokes system (1.1)–(1.2). Following [3], we first imbed the system (2.19)–(2.22) in a family of autonomous systems depending on a parameter $\sigma \in \mathbb{T}^N$, taking $f(\Phi(x,t),t) = \tilde{f}(\Phi(x,t), \alpha t + \sigma)$:

$$v_t = u_x, \tag{3.17}$$

$$u_t + (A^2 v^{-1})_x = \left(\frac{u_x}{v}\right)_x + \tilde{f}\left(\int_0^x v(\xi,t)d\xi, \omega\right), \tag{3.18}$$

$$\frac{d\omega(t)}{dt} = \alpha, \tag{3.19}$$

$$u(0,t) = u(1,t) = 0, \quad \int_0^1 v(x,t)dx = 1, \quad t \ge 0, \tag{3.20}$$

$$u(x,0) = u_0, \quad v(x,0) = v_0, \quad \text{and} \quad \omega(0) = \sigma. \tag{3.21}$$

The existence and regularity theory of section 2 applies to the system (3.17)–(3.21), and all the estimates of Lemmas 2.4 and 2.5 hold uniformly for $\sigma \in \mathbb{T}^N$.

To construct the decomposition described in Proposition 3.8, we let $(\bar{v}, \bar{u}, \bar{\omega})$ and (v, u, ω) be solutions of (3.17)–(3.21) with respective initial values $(\bar{v}_0, \bar{u}_0, \bar{\sigma})$ and (v_0, u_0, σ), and we set

$$\varphi = v - \bar{v}, \quad \psi = u - \bar{u}, \quad \eta = \omega - \bar{\omega}, \tag{3.22}$$

$$\Phi(x,t) = \int_0^x v(s,t)\,ds, \quad \bar{\Phi}(x,t) = \int_0^1 \bar{v}(s,t)\,ds, \quad \text{and} \tag{3.23}$$

$$R(x,t) = \Phi(x,t) - \bar{\Phi}(x,t) = \int_0^x \varphi(\xi,t)d\xi. \tag{3.24}$$

Then

$$R(0,t) = R(1,t) = 0 \tag{3.25}$$

and

$$\text{(3.26)}\quad \begin{cases} \varphi_t - \psi_x = 0, \\ \psi_t + (q\varphi)_x = \varepsilon\psi_{xx} + (\zeta\varphi)_{xt} + hR + \bar{h}\eta, \\ \dfrac{d\eta}{dt} = 0, \\ \varphi_0(x,0) = v_0(x) - \bar{v}_0(x), \quad \psi_0(x,0) = u_0(x) - \bar{u}_0(x), \quad \eta(0) = \sigma - \bar{\sigma}, \end{cases}$$

where

$$\text{(3.27)}\quad \Delta\tilde{f} = \tilde{f}(\Phi(x,t), \alpha t + \sigma) - \tilde{f}(\bar{\Phi}(x,t), \alpha t + \bar{\sigma}) = h(x,t)R(x,t) + \bar{h}(x,t)\eta(t),$$

$$\text{(3.28)}\quad p(v) - p(\bar{v}) = q(x,t)\varphi(x,t), \quad \text{with} \quad q(x,t) = -\frac{A^2}{v(x,t)\bar{v}(x,t)},$$

$$\frac{u_x}{v} - \frac{\bar{u}_x}{v} = \varepsilon\psi_x + (\log v - \varepsilon v)_t - (\log\bar{v} - \varepsilon\bar{v})_t = \varepsilon\psi_x + (\zeta\varphi)_t,$$

where

$$\text{(3.29)}\qquad \zeta = \zeta(x,t) = \frac{1}{\tilde{v}(x,t)} - \varepsilon \quad \text{for some} \quad \tilde{v}(x,t) \in [v(x,t), \bar{v}(x,t)],$$

and ε is defined by $\varepsilon = \frac{1}{2}\inf[v(x,t)^{-1}],\ x \in (0,1),\ t \geq T_0 + 1$.

We now let λ be a large positive constant and k a large positive integer, to be chosen later, and introduce the following decomposition of (φ, ψ, η):

$$(\varphi, \psi, \eta) = (\varphi^I, \psi^I, 0) + (\varphi^{II}, \psi^{II}, \eta),$$

where (φ^I, ψ^I) and $(\varphi^{II}, \psi^{II})$ are the solutions of

$$\text{(3.30)}\quad \begin{cases} \varphi^I_t - \psi^I_x = 0, \\ \psi^I_t + (q\varphi^I)_x = \varepsilon\psi^I_{xx} + (\zeta\varphi^I)_{xt} + \lambda\tilde{v}(x,t)(P_k\psi^I_{xx}) + hQ_kR^I, \\ \varphi^I(x,0) = \varphi_0(x,0), \quad \psi^I(x,0) = \psi_0(x), \end{cases}$$

$$\text{(3.31)}\quad \begin{cases} \varphi^{II}_t - \psi^{II}_x = 0, \\ \psi^{II}_t + (q\varphi^{II})_x = \varepsilon\psi^{II}_{xx} + (\zeta\varphi^{II})_{xt} - \lambda\tilde{v}(P_k\psi^I_{xx}) + hP_kR^I + hR^{II} + \bar{h}\eta, \\ \varphi^{II}(x,0) = \psi^{II}(x,0) = 0, \end{cases}$$

where

$$\text{(3.32)}\qquad P_k\psi = \sum_{|j|\leq k} \widehat{\psi}_j e^{2\pi ijx} \quad \text{and} \quad Q_k\psi = \psi - P_k\psi.$$

It is readily verified that the sum of the solutions $(\varphi^I, \psi^I, 0)$ and $(\varphi^{II}, \psi^{II}, \eta)$ of the above systems is indeed the solution (φ, ψ, η) of (3.17)–(3.21).

The proof that this decomposition satisfies the hypotheses of Proposition 3.8 will depend on certain a priori estimates for solutions of the systems (3.30) and (3.31). We state these estimates without proof below in Theorem 3.9, and we then complete the proof of Theorem 1.1 by showing how these a priori estimates can be used to fit the Navier–Stokes system (2.19)–(2.22) into the abstract framework of section 3.1. The proof of Theorem 3.9 will be deferred to section 4.

THEOREM 3.9. *For k, λ, and k/λ sufficiently large, there exist positive constants $r \in (0,1)$, ν, ν_1, ν_2, ν_3, and K_0, all depending only on A and C_Σ (see* (2.1) *and*

Theorem 2.1*), such that if* (v_0, u_0) *and* $(\bar v_0, \bar u_0)$ *are in the uniform attractor* $\mathcal{A}_\Sigma$*, then*

$$|\varphi_x^I(\cdot,t)|^2 + |\psi_x^I(\cdot,t)|^2 \le K_0 e^{-\nu t}\big[|\varphi_{0x}^I| + |\psi_{0x}^I|^2\big], \tag{3.33}$$

$$|\varphi^I(\cdot,t)|^2 + |\psi^I(\cdot,t)|^2 \le K_0 e^{\nu_1 t}\big[|\varphi_0^I|^2 + |\psi_0^I|^2\big], \tag{3.34}$$

$$|\varphi^I(\cdot,t)|_{H^r}^2 + |\psi^I(\cdot,t)|_{H^r}^2 \le K_0 e^{-\nu_2 t}\big[|\varphi_0^I|_{H^r}^2 + |\psi_0^I|_{H^r}^2\big], \tag{3.35}$$

$$|\varphi_x^{II}(\cdot,t)|^2 + |\psi_x^{II}(\cdot,t)|^2 \le K_0 e^{\nu_3 t}\big[|\varphi_0^I|_{H^r}^2 + |\psi_0^I|_{H^r}^2\big]. \tag{3.36}$$

Proof of Theorem 1.1. We first show that the global attractor $\mathcal{A}$ of Theorem 2.3 has finite fractal dimension by applying the result of Lemma 1.2. To do this, we let H be the Hilbert space $H = H^r \times H_0^r \times \mathbb{T}^N$, where r is as above in Theorem 3.9, and we take $M = \mathcal{A}$, which is compact in H. To check the hypotheses (1.5) and (1.6) we let $w_1 = (v_0, u_0, \sigma_0)$ and $w_2 = (\bar v_0, \bar u_0, \bar\sigma_0)$ be two points in $\mathcal{A}$ and write $S(t)w_1 - S(t)w_2 = W^I(t) + W^{II}(t)$, with $W^I(t) = (\varphi^I(t), \psi^I(t), 0)$ and $W^{II}(t) = (\varphi^{II}(t), \psi^{II}(t), \eta(t))$, where $\eta(t) = \alpha t + \sigma - \bar\sigma$, $(\varphi^I(t), \psi^I(t))$ is the solution of (3.30) with initial data $(v_0 - \bar v_0, u_0 - \bar u_0)$, and $(\varphi^{II}(t), \psi^{II}(t))$ is the solution of (3.31) with zero initial data. Then for t^* large,

$$|W^I(t^*)|_{H^r\times H_0^r\times\mathbb{T}^N} \le 2^{-1/2}\delta |w_1 - w_2|_{H^r\times H_0^r\times\mathbb{T}^N} \tag{3.37}$$

with $\delta < 1$, by (3.35). Next, if $\tilde P_n$ is the orthogonal projector onto the span of $\{e^{2k\pi i x}\}_{|k|\le n}$, and if $\tilde Q_n = I - \tilde P_n$, then

$$|\tilde Q_n g|_{H^r} \le n^{1-r}|g|_{H^1}$$

for $g \in H^1([0,1])$. Letting P_n be the projector $P_n = (\tilde P_n, \tilde P_n, Id)$ on $L^2 \times L^2 \times \mathbb{T}^N$ and $Q_n = I - P_n$, we then have from (3.36) that, for a generic constant C,

$$\begin{aligned} |Q_n W^{II}(t^*)|_{H^r\times H_0^r\times\mathbb{T}^N} &\le Cn^{1-r}|(\phi^{II}(t^*), \psi^{II}(t^*), 0)|_{H^1\times H_0^1\times\mathbb{T}^N} \\ &\le Cn^{1-r}e^{\nu_3 t^*}|w_1 - w_2|_{H^r\times H_0^r\times\mathbb{T}^N} \\ &\le 2^{-1/2}\delta |w_1 - w_2|_{H^r\times H_0^r\times\mathbb{T}^N} \end{aligned} \tag{3.38}$$

if n is chosen sufficiently large, depending on t^*. Taking the square root of the sum of the squares of (3.37) and (3.38), we then obtain (1.6) for $S = S(t^*)$. The Lipschitz continuity (1.5) then follows immediately from (3.35) and (3.36). Since $\mathcal{A}$ is compact, by Theorem 2.3, the result of Lemma 1.2 applies to show that $\mathcal{A}$ has finite fractal dimension in $H^r \times H_0^r \times \mathbb{T}^N$, and therefore that its projection $\mathcal{A}_\Sigma$ has finite fractal dimension in $H^r \times H_0^r$, as asserted in Theorem 1.1.

To prove the existence of a finite-dimensional exponential attractor for the semigroup $S(t)$, we apply the results of Theorem 3.7 and Proposition 3.8 as follows. Let H and V be the Hilbert spaces $H = H^r \times H_0^r \times \mathbb{T}^N$, where r is as above in Theorem 3.9, and $V = H^1 \times H_0^1 \times \mathbb{T}^N$, and let $\mathcal{Y}$ be the absorbing ball $\mathcal{B}$ of Theorem 2.3, which is compact in H. Taking P_n, Q_n, W^I, and W^{II} exactly as above, we then obtain (3.15) and (3.16) directly from the conclusions of Theorem 3.9. Proposition 3.8 and Theorem 3.7 then apply to show that the semigroup $S(t)$ satisfies the squeezing property, and that therefore there exists an exponential attractor $\mathcal{M}$ of finite fractal dimension. The uniform exponential attractor $\mathcal{M}_\Sigma$ for the process $\mathcal{U}_\sigma$ is then obtained as the projection of $\mathcal{M}$ onto $H^r \times H_0^r$. This completes the proof of Theorem 1.1. ∎

4. Proof of Theorem 3.9. This section is devoted to the proof of Theorem 3.9. We will concentrate on the proof of the inequalities (3.33) and (3.34) and then obtain the inequality (3.35) (which corresponds to (3.15) in Proposition 3.8) by the Riesz–Thorin interpolation theorem. The last inequality (3.36) (which corresponds to the inequality (3.16) in Proposition 3.8) is straightforward and will be sketched.

We assume throughout this section that all the hypotheses and notation of Theorem 3.9 are in force, and we begin with the following technical lemma.

LEMMA 4.1. *Assume that the initial data* (v_0, u_0), $(\bar v_0, \bar u_0)$ *are in the attractor* $\mathcal{A}_\Sigma$ *and let* $(v(t), u(t))$, $(\bar v(t), \bar u(t))$ *be the corresponding solutions of the Navier–Stokes equations, so that, as a consequence of Lemmas* 2.4 *and* 2.5,

$$0 < -q(x,t) \le M_0, \quad |\zeta(\cdot,t)|_{L^\infty} \le M_0, \quad |\zeta_t(\cdot,t)|_{L^\infty} \le M_0, \tag{4.1}$$

$$|\zeta_x(\cdot,t)|_{L^2} \le M_0, \quad |\zeta_{xt}(\cdot,t)|_{L^\infty} \le M_0, \quad \frac{1}{M_0} \le \tilde v(x,t) \le M_0, \tag{4.2}$$

where M_0 *is a constant depending only on* A *and* C_Σ. *Then there exists a numerical constant* C *such that*

$$|\psi^I_x|^2_{L^\infty} \le C M_0^8 \left[\frac{1}{\lambda^2}|\varphi^I|^2 + |\psi^I_x|^2 + \frac{1}{k}|\psi^I_{xx}|^2 + \frac{1}{\lambda^2}|\psi^I_t|^2 + \frac{1}{\lambda^2}|\varphi^I||\varphi^I_x| \right] \tag{4.3}$$

and

$$|\psi^I_{xx}|^2 \le C M_0^8 \left[|\varphi^I|^2 + M_0^4 |\varphi^I_x|^2 + |\psi^I_x|^2 + |\psi^I_t|^2 \right]. \tag{4.4}$$

Proof. Let $w = \frac{\psi^I_x}{\tilde v} + (\zeta_t - q)\varphi^I + \lambda \tilde v P_k \psi^I_x$. Thanks to Lemmas 2.4 and 2.5

$$|w|^2 \le 2M_0^2(|\psi^I_x|^2 + |\varphi^I|^2 + \lambda^2 |P_k \psi^I_x|^2) \le C M_0^2 \big[\lambda^2 |\psi^I_x|^2 + |\varphi^I|^2 \big], \tag{4.5}$$

where C denotes a generic numerical constant. We note that equations (3.30) imply that $w_x = \psi^I_t + \lambda \tilde v_x P_k \psi^I_x - h Q_k R^I$. Therefore,

$$|w_x| \le |\psi^I_t| + \lambda |\tilde v_x| |P_k \psi^I_x|_{L^\infty} + |h|_{L^\infty} |Q_k R^I|. \tag{4.6}$$

Since $|Q_k R^I| \le |\varphi^I|$, Lemmas 2.4 and 2.5 yield

$$|w_x|^2 \le C M_0^2 \big(|\psi^I_t|^2 + |\varphi^I|^2 + \lambda^2 |P_k \psi_x|^2_{L^\infty} \big). \tag{4.7}$$

Now using Agmon's inequalities

$$|P_k \psi^I_x|^2_{L^\infty} \le |\psi^I_x|^2_{L^\infty} + \frac{1}{k}|\psi^I_{xx}|^2, \tag{4.8}$$

$$|w|^2_{L^\infty} \le |w|^2 + 2|w||w_x|, \tag{4.9}$$

we obtain

$$|w|^2_{L^\infty} \le C M_0^2 \left[\lambda^2 |\psi^I_x|^2 + |\varphi^I|^2 + \frac{\lambda^2}{k}|\psi^I_{xx}|^2 + \big(\lambda |\psi^I_x| + |\varphi^I| \big) \big(|\psi^I_t| + \lambda |\psi^I_x|_{L^\infty} \big) \right]. \tag{4.10}$$

Furthermore, from the definition of w, we can write

$$\psi^I_x = \left(\frac{1}{\tilde v} + \lambda \tilde v \right)^{-1} [w - (\zeta_t - q)\varphi^I + \lambda \tilde v Q_k \psi^I_x], \tag{4.11}$$

which implies that, for $\lambda \geq 1$,

$$|\psi_x^I|_{L^\infty}^2 \leq \frac{CM_0^2}{\lambda^2}\left[|w|_{L^\infty}^2 + |\zeta_t - q|_{L^\infty}^2|\varphi^I|_{L^\infty}^2 + \lambda^2|\tilde{v}|_{L^\infty}^2|Q_k\psi_x^I|_{L^\infty}^2\right], \tag{4.12}$$

and since $|Q_k\psi_x^I|_{L^\infty}^2 \leq \frac{1}{k^2}|\psi_{xx}^I|^2$, we obtain, thanks to Lemmas 2.4 and 2.5,

$$\begin{aligned}|\psi_x^I|_{L^\infty}^2 \leq CM_0^4 &\left[|\psi_x^I|^2 + \frac{1}{\lambda^2}|\varphi^I|^2 + \frac{1}{k}|\psi_{xx}^I|^2 + \left(\frac{1}{\lambda}|\psi_x^I| + \frac{1}{\lambda^2}|\varphi^I|\right)\left(|\psi_t^I| + \lambda|\psi_x^I|_{L^\infty}\right)\right] \\ &+ \frac{CM_0^4}{\lambda^2}|\varphi^I|_{L^\infty}^2 + \frac{CM_0^4}{k^2}|\psi_{xx}^I|^2,\end{aligned} \tag{4.13}$$

and with the inequality

$$CM_0^4\left(|\psi_x^I| + \frac{1}{\lambda}|\varphi|\right)|\psi_x^I|_{L^\infty} \leq \frac{1}{2}|\psi_x^I|_{L^\infty}^2 + CM_0^8|\psi_x^I|^2 + \frac{CM_0^8}{\lambda^2}|\varphi^I|^2, \tag{4.14}$$

we conclude that

$$|\psi_x^I|_{L^\infty}^2 \leq CM_0^8\left(|\psi_x^I|^2 + \frac{1}{\lambda^2}|\varphi^I|^2 + \frac{1}{k}|\psi_{xx}^I|^2 + \frac{1}{\lambda^2}|\psi_t^I|^2\right) + \frac{CM_0^4}{\lambda^2}|\varphi^I||\varphi_x^I|. \tag{4.15}$$

This completes the proof of (4.3).

In order to prove inequality (4.4), we rewrite the equations (3.30) in the form

$$\psi_t^I + q_x\varphi^I + q\varphi_x^I = \frac{1}{\tilde{v}}\psi_{xx}^I + \zeta_x\psi_x^I + \zeta_t\varphi_x^I + \zeta_{tx}\varphi^I + \lambda\tilde{v}P_k\psi_{xx}^I + hQ_kR^I. \tag{4.16}$$

Let $\mathcal{F}$ be defined by

$$\mathcal{F} = \psi_t^I + q_x\varphi^I + q\varphi_x^I - \zeta_x\psi_x^I - \zeta_t\varphi_x^I - \zeta_{tx}\varphi^I - hQ_kR^I. \tag{4.17}$$

We have

$$\frac{1}{\tilde{v}}P_k\psi_{xx}^I + \frac{1}{\tilde{v}}Q_k\psi_{xx}^I + \lambda\tilde{v}P_k\psi_{xx}^I = \mathcal{F}, \tag{4.18}$$

and, with Lemmas 2.4 and 2.5,

$$\begin{aligned}|\mathcal{F}| &\leq |\psi_t^I| + |q_x||\varphi^I|_{L^\infty} + |q|_{L^\infty}|\varphi_x^I| + |\zeta_x||\psi_x^I|_{L^\infty} \\ &\quad + |\zeta_t|_{L^\infty}|\varphi_x^I| + |\zeta_{tx}||\varphi^I|_{L^\infty} + |h|_{L^\infty}|Q_kR^I| \\ &\leq CM_0\left[|\psi_t| + |\varphi_x^I| + |\varphi^I| + |\psi_x^I|_{L^\infty}\right].\end{aligned} \tag{4.19}$$

We multiply (4.18) by $\frac{Q_k\psi_{xx}}{\tilde{v}}$ and integrate with respect to x to obtain

$$\begin{aligned}|\tilde{v}^{-1}Q_k\psi_{xx}^I|^2 &= \int_0^1 \mathcal{F}\tilde{v}^{-1}Q_k\psi_{xx}^I\,dx \\ &\quad - \int_0^1 \tilde{v}^{-2}Q_k\psi_{xx}^IP_k\psi_{xx}^I\,dx \leq \left[|\mathcal{F}| + |\tilde{v}^{-1}P_k\psi_{xx}^I|\right]|\tilde{v}^{-1}Q_k\psi_{xx}^I|,\end{aligned}$$

so that

$$|\tilde{v}^{-1}Q_k\psi_{xx}^I|^2 \leq C|\mathcal{F}|^2 + |\tilde{v}^{-1}P_k\psi_{xx}^I|^2.$$

Hence

$$|Q_k\psi^I_{xx}|^2 \le CM_0^2[|\mathcal{F}|^2 + |P_k\psi^I_{xx}|^2] \le CM_0^4\Big[|\psi_t|^2 + |\varphi^I_x|^2 + |\varphi^I|^2 + |\psi^I_x|^2_{L^\infty}\Big] + |P_k\psi^I_{xx}|^2. \tag{4.20}$$

Now we multiply (4.18) by $\tilde v P_k\psi^I_{xx}$ and obtain

$$\int_0^1 (1+\lambda\tilde v^2)(P_k\psi^I_{xx})^2dx = \int_0^1 \mathcal{F}\tilde v P_k\psi_{xx}\,dx \le M_0|\mathcal{F}||P_k\psi_{xx}| \le CM_0^2|\mathcal{F}|^2 + \frac{1}{2}|P_k\psi_{xx}|^2. \tag{4.21}$$

Therefore

$$|P_k\psi^I_{xx}|^2 \le \frac{CM_0^4}{\lambda}|\mathcal{F}|^2 \le \frac{CM_0^4}{\lambda}\Big[|\psi_t|^2 + |\varphi^I_x|^2 + |\varphi^I|^2 + |\psi^I_x|^2_{L^\infty}\Big]. \tag{4.22}$$

Combining (4.20) and (4.22), we obtain

$$|\psi^I_{xx}|^2 \le CM_0^4\Big[|\psi^I_t|^2 + |\varphi^I|^2 + |\varphi^I_x|^2 + |\psi^I_x|^2_{L^\infty}\Big]. \tag{4.23}$$

Finally, applying Agmon's inequality, we obtain

$$|\psi^I_{xx}|^2 \le CM_0^8\Big[|\psi^I_t|^2 + |\varphi^I|^2 + |\varphi^I_x|^2 + |\psi^I_x|^2\Big]. \tag{4.24}$$

This concludes the proof of Lemma 4.1. □

Proof of Theorem 3.9.

***L*²-estimate for ψ^I.** We rewrite (3.30) in the form

$$\psi^I_t + (q\varphi^I)_x = \left(\frac{\psi^I_x}{\tilde v} + \zeta_t\varphi^I\right)_x + \lambda\tilde v P_k\psi^I_{xx} + hQ_kR^I \tag{4.25}$$

and multiply by $\frac{\psi^I}{\tilde v}$ to obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_0^1 \frac{(\psi^I)^2}{\tilde v}dx - \frac{1}{2}\int_0^1 (\psi^I)^2\zeta_t\,dx &= \int_0^1 q\varphi^I\left(\frac{\psi^I_x}{\tilde v} + \psi^I\zeta_x\right)dx\\ &\quad - \int_0^1\left(\frac{\psi^I_x}{\tilde v} + \zeta_t\varphi^I\right)\left(\frac{\psi^I_x}{\tilde v} + \psi^I\zeta_x\right)dx\\ &\quad - \lambda\int_0^1 (P_k\psi^I_x)^2dx + \int_0^1 \frac{h}{\tilde v}\psi^I Q_kR^I\,dx.\end{aligned} \tag{4.26}$$

Since $|\psi^I| \le |\psi^I_x|$, Lemmas 2.4 and 2.5 yield

$$\frac{1}{2}\frac{d}{dt}\int_0^1 \frac{(\psi^I)^2}{\tilde v}dx + \int_0^1\left(\frac{(\psi^I_x)^2}{\tilde v^2} + \lambda(P_k\psi^I_x)^2\right)dx \le CM_0^2\big[|\varphi^I||\psi^I_x| + |\psi^I_x|^2\big], \tag{4.27}$$

and since

$$\lambda|P_k\psi^I_x|^2 = \lambda|\psi^I_x|^2 - \lambda|Q_k\psi^I_x|^2 \ge \lambda|\psi^I_x|^2 - \frac{\lambda}{k^2}|\psi^I_{xx}|^2 \tag{4.28}$$

we obtain

$$\begin{aligned}\frac{d}{dt}\int_0^1 \frac{(\psi^I)^2}{2\tilde v}dx + \lambda|\psi_x^I|^2 &\le CM_0^2|\varphi^I||\psi_x^I| + CM_0^2|\psi_x^I|^2 + C\frac{\lambda}{k^2}|\psi_{xx}^I|^2\\ &\le \frac{\lambda}{2}|\psi_x^I|^2 + C\frac{M_0^4}{\lambda}|\varphi^I|^2 + CM_0^2|\psi_x^I|^2\\ &\quad + C\frac{\lambda}{k^2}M_0^8\Big[|\varphi^I|^2 + |\varphi_x^I|^2 + |\psi_x^I|^2 + |\psi_t^I|^2\Big].\end{aligned} \tag{4.29}$$

Now assuming that

$$\lambda \ge 4CM_0^2 \quad \text{and} \quad k^2 \ge 4CM_0^8, \tag{4.30}$$

we obtain

$$\frac{d}{dt}\int_0^1 \frac{(\psi^I)^2}{\tilde v}dx + \frac{\lambda}{2}|\psi_x^I|^2 \le CM_0^8\left(\frac{1}{\lambda}+\frac{\lambda}{k^2}\right)|\varphi^I|^2 + CM_0^8\frac{\lambda}{k^2}\big(|\varphi_x^I|^2 + |\psi_t^I|^2\big). \tag{4.31}$$

***H^1*-estimate for ψ^I.** We multiply (4.25) by $\frac{\psi_t^I}{\tilde v}$ and integrate to obtain

$$\begin{aligned}\int_0^1 \frac{(\psi_t^I)^2}{\tilde v}dx + \int_0^1 \frac{\psi_t^I}{\tilde v}(q\varphi^I)_x dx &= -\int_0^1 \left(\frac{\psi_t^I}{\tilde v}\right)_x\left(\frac{\psi_x^I}{\tilde v} + \zeta_t\varphi^I\right)dx\\ &\quad - \frac{\lambda}{2}\frac{d}{dt}\int_0^1 (P_k\psi_x^I)^2dx + \int_0^1 \frac{h}{\tilde v}\psi_t^I Q_k R^I\ dx.\end{aligned}$$

We note that

$$\begin{aligned}\int_0^1 \frac{\psi_t^I}{\tilde v}(q\varphi^I)_x dx &= -\frac{d}{dt}\int_0^1 \frac{q}{\tilde v}\psi_x^I\varphi^I\ dx + \int_0^1 (\psi_x^I)^2\frac{q}{\tilde v}dx + \int_0^1 \frac{q_t}{\tilde v}\varphi^I\psi_x^I\ dx\\ &\quad + \int_0^1 \psi_x^I\varphi^I q\zeta_t\,dx - \int_0^1 \zeta_x\psi_t^I\tilde v_x q\varphi^I\ dx + \int_0^1 \frac{\psi_t^I}{\tilde v}q\varphi_x^I\ dx,\end{aligned} \tag{4.32}$$

which implies

$$\int_0^1 \frac{\psi_t^I}{\tilde v}(q\varphi^I)_x dx = -\frac{d}{dt}\int_0^1 \frac{q}{\tilde v}\psi_x^I\varphi^I\ dx + \mathcal{E}_1, \tag{4.33}$$

with

$$|\mathcal{E}_1| \le CM_0^2\left[|\psi_x^I|^2 + |\psi_x^I||\varphi^I| + |\psi_t^I||\varphi^I| + |\psi_t^I||\varphi^I|\right]. \tag{4.34}$$

Furthermore,

$$\begin{aligned}\int_0^1 \left(\frac{\psi_t^I}{\tilde v}\right)_x\left(\frac{\psi_x^I}{\tilde v} + \zeta_t\varphi^I\right)dx &= \frac{d}{dt}\int_0^1 \left(\frac{(\psi_x^I)^2}{2\tilde v^2} + \frac{\zeta_t}{\tilde v}\psi_x^I\varphi^I\right)dx + \int_0^1 \frac{(\psi_x^I)^2\tilde v_t}{\tilde v^3}dx\\ &\quad - \int_0^1 \frac{\zeta_{tt}}{\tilde v}\psi_x^I\varphi^I\ dx - \int_0^1 \frac{\zeta_t(\psi_x^I)^2}{\tilde v}dx\\ &\quad + \int_0^1 \left[\frac{\zeta_x}{\tilde v}\psi_t^I\psi_x^I + \zeta_x\zeta_t\psi_t^I\varphi^I\right]dx.\end{aligned} \tag{4.35}$$

Hence

$$\int_0^1 \left(\frac{\psi_t^I}{\tilde v}\right)_x \left(\frac{\psi_x^I}{\tilde v} - \zeta_t \varphi^I\right) dx = \frac{d}{dt}\int_0^1 \left[\frac{(\psi_x^I)^2}{2\tilde v^2} + \frac{\zeta_t \psi_x^I \varphi^I}{\tilde v}\right] dx + \mathcal{E}_2, \tag{4.36}$$

with

$$|\mathcal{E}_2| \le CM_0^2\big[|\psi_x^I|^2 + |\varphi^I|^2\big] + CM_0^2\big[|\psi_t^I||\psi_x^I| + |\psi_t^I||\varphi^I|^{1/2}|\varphi_x^I|^{1/2}\big]. \tag{4.37}$$

Therefore

$$\begin{aligned}
&\frac{d}{dt}\int_0^1 \left[\frac{(\psi_x^I)^2}{2\tilde v^2} + \frac{\zeta_t \psi_x^I \varphi^I}{\tilde v} - \frac{q}{\tilde v}\psi_x^I \varphi^I + \frac{\lambda}{2}(P_k \psi_x^I)^2\right] dx + \int_0^1 \frac{(\psi_t^I)^2}{\tilde v} dx \\
&\quad \le CM_0^2\big[|\psi_x^I|^2 + |\varphi^I|^2\big] + CM_0^2|\psi_t^I|\big[|\varphi^I| + |\psi_x^I| + |\varphi_x^I|^{1/2}|\varphi^I|^{1/2}\big] \\
&\quad \le CM_0^2\big[|\psi_x^I|^2 + |\varphi^I|^2\big] + \frac{1}{2M_0}|\psi_t^I|^2 + CM_0^5\big[|\psi_x^I|^2 + |\varphi^I|^2\big] + CM_0^5|\varphi^I||\varphi_x^I|
\end{aligned} \tag{4.38}$$

and

$$\begin{aligned}
\frac{d}{dt}\int_0^1 \left[\frac{(\psi_x^I)^2}{2\tilde v} + \frac{\lambda}{2}(P_k\psi_x^I)^2 + \frac{\zeta_t - q}{\tilde v}\psi_x^I\varphi^I\right] dx + \frac{1}{2M_0}|\psi_t^I|^2 \le CM_0^5\big[|\psi_x^I|^2 + |\varphi^I|^2\big] \\
+ CM_0^5|\varphi^I||\varphi_x^I|.
\end{aligned} \tag{4.39}$$

***L^2*-estimate for φ^I.** We write (3.30) in the form

$$\psi_t^I + (q\varphi^I)_x = \left(\frac{1}{\tilde v}\varphi_t^I + \zeta_t \varphi^I\right)_x + \lambda \tilde v P_k \psi_{xx}^I + hQ_k R^I, \tag{4.40}$$

and noting that

$$R_x^I = \varphi^I \quad \text{and} \quad R_t^I = \psi^I, \tag{4.41}$$

we multiply (4.40) by R^I and integrate to obtain

$$\begin{aligned}
\frac{d}{dt}\int_0^1 R^I\psi^I\,dx - \int_0^1 (\psi^I)^2 dx + \int_0^1 (-q)(\varphi^I)^2\,dx = -\int_0^1 \left[\frac{1}{\tilde v}\varphi^I\varphi_t^I + \zeta_t(\varphi^I)^2\right] dx \\
- \lambda\int_0^1 (\tilde v R^I)_x P_k\psi_x^I\,dx + \int_0^1 hR^I Q_k R^I\,dx.
\end{aligned}$$

Since

$$\frac{d}{dt}\int_0^1 \frac{1}{\tilde v}(\varphi^I)^2 dx = 2\int_0^1 \frac{1}{\tilde v}\varphi^I\varphi_t^I\,dx + \int_0^1 \zeta_t(\varphi^I)^2 dx, \tag{4.42}$$

we can write

$$\begin{aligned}
&\frac{d}{dt}\int_0^1 \left[R^I\psi^I + \frac{(\varphi^I)^2}{\tilde v}\right] dx + \int_0^1 (-q)(\varphi^I)^2 dx = \int_0^1 |\psi^I|^2 dx + \int_0^1 \frac{1}{\tilde v}\varphi^I\psi_x^I\,dx \\
&\qquad - \lambda\int_0^1 (R^I\tilde v)_x P_k\psi_x^I\,dx + \int_0^1 hR^I Q_k R^I\,dx \\
&\qquad \le C|\psi^I|^2 + M_0|\varphi^I||\psi_x^I| + \lambda M_0|\varphi^I||P_k\psi_x^I| + C\frac{M_0}{k}|\varphi^I|^2,
\end{aligned} \tag{4.43}$$

and since $M_0^{-1} \le -q(x,t)$ for $x \in (0,1)$ and $t \ge 0$, and $|\psi^I| \le |\psi_x^I|$, we conclude that for $k \ge 4CM_0^2$,

$$\frac{d}{dt}\int_0^1 \left[R^I\psi^I + \frac{(\varphi^I)^2}{\tilde v}\right] dx + \frac{1}{2M_0}|\varphi^I|^2 \le CM_0^3\lambda^2|\psi_x^I|^2. \tag{4.44}$$

H^1-estimate for φ^I. We write the equation satisfied by ψ^I in the form

$$\psi_t^I + (q_x\varphi^I + q\varphi_x^I) = \left(\frac{\varphi_x^I}{\tilde v}\right)_t + (\zeta_x\varphi^I)_t + \lambda\tilde v P_k\varphi_{xt} + hQ_kR^I \tag{4.45}$$

and multiply by $\frac{\varphi_x}{\tilde v}$ and integrate to obtain

$$\begin{aligned}\int_0^1 \frac{\varphi_x^I}{\tilde v}\psi_t^I\,dx + \int_0^1 \frac{\varphi_x^I}{\tilde v}(q_x\varphi^I + q\varphi_x^I)dx &= \int_0^1 \frac{\varphi_x^I}{\tilde v}\left(\frac{\varphi_x^I}{\tilde v}\right)_t dx \\ &+ \int_0^1 (\zeta_x\varphi)_t\frac{\varphi_x}{\tilde v}\,dx + \lambda\int \frac{\varphi_x^I}{\tilde v}\tilde v P_k\varphi_{xt}^I\,dx + \int_0^1 \frac{\varphi_x^I}{\tilde v}hQ_kR^I\,dx.\end{aligned} \tag{4.46}$$

Therefore, using Lemmas 2.4 and 2.5,

$$\begin{aligned}\frac{d}{dt}\int_0^1 \left(\frac{(\varphi_x^I)^2}{2\tilde v^2} + \frac{\lambda}{2}(P_k\varphi_x^I)^2\right)dx + \frac{1}{M_0^2}|\varphi_x^I|^2dx &\le CM_0|\varphi_x^I||\psi_t^I| + CM_0^2|\varphi^I|^{1/2}|\varphi_x^I|^{3/2} \\ &+ CM_0^2|\varphi_x^I||\varphi^I| + CM_0^2|\psi_x|_{L^\infty}|\varphi_x^I|.\end{aligned} \tag{4.47}$$

Hence, using (4.4),

$$\frac{d}{dt}\int_0^1 \left(\frac{(\varphi_x^I)^2}{2\tilde v^2} + \frac{\lambda}{2}(P_k\varphi_x^I)^2\right)dx + \frac{1}{2M_0^2}|\varphi_x^I|^2 \le CM_0^{20}\big[|\psi_t^I|^2 + |\varphi^I|^2 + |\psi_x^I|^2\big]. \tag{4.48}$$

Let γ_1, γ_2, and γ_3 be large positive numbers, to be determined below, and multiply (4.39) by γ_1, (4.44) by γ_2, and (4.31) by γ_3. The resulting inequalities and inequality (4.48) are

$$\frac{d}{dt}\int_0^1 \left(\frac{(\varphi_x^I)^2}{2\tilde v^2} + \frac{\lambda}{2}(P_k\varphi_x^I)^2\right)dx + \frac{1}{2M_0^2}|\varphi_x^I|^2 \le CM_0^{20}\big[|\psi_t^I|^2 + |\varphi^I|^2 + |\psi_x^I|^2\big], \tag{4.49}$$

$$\gamma_2\frac{d}{dt}\int_0^1 \left[R^I\psi^I + \frac{(\varphi^I)^2}{\tilde v}\right]dx + \frac{\gamma_2}{2M_0}|\varphi^I|^2 \le C\gamma_2\lambda^2M_0^3|\psi_x^I|^2, \tag{4.50}$$

$$\begin{aligned}\gamma_1\frac{d}{dt}\int_0^1 \left[\frac{(\psi_x^I)^2}{2\tilde v} + \frac{\lambda}{2}(P_k\psi_x^I)^2 + \frac{\zeta_t - q}{\tilde v}\psi_x^I\varphi^I\right]dx + \frac{\gamma_1}{2M_0}|\psi_t^I|^2 \\ \le \gamma_1CM_0^5\big[|\psi_x^I|^2 + |\varphi^I|^2\big] + \gamma_1CM_0^5|\varphi^I||\varphi_x^I| \\ \le CM_0^5\big[|\psi_x^I|^2 + |\varphi^I|^2\big] + \frac{1}{2M_0^2}|\varphi_x|^2 + C\gamma_1^2M_0^{12}|\varphi^I|^2,\end{aligned} \tag{4.51}$$

$$\gamma_3\frac{d}{dt}\int_0^1 \frac{(\psi^I)^2}{\tilde v}dx + \gamma_3\frac{\lambda}{2}|\psi_x^I|^2 \le C\gamma_3M_0^8\left(\frac{1}{\lambda} + \frac{\lambda}{k^2}\right)|\varphi^I|^2 + C\gamma_3M_0^8\frac{\lambda}{k^2}\big(|\varphi_x^I|^2 + |\psi_t^I|^2\big). \tag{4.52}$$

We choose γ_1 and γ_2 to be large enough so that

$$\gamma_1 \geq 4CM_0^{21}, \quad \gamma_2 \geq 8CM_0^{21}, \quad \gamma_1 \leq C\frac{\gamma_2}{M_0^6}, \quad \gamma_1^2 \leq C\frac{\gamma_2}{M_0^{13}}. \tag{4.53}$$

Then choose γ_3 large enough so that $\gamma_3\lambda/2 > 2C\gamma_2\Lambda^2 M_0^3$, and finally choose λ and k large enough so that

$$\lambda \geq 4CM_0^{20}, \quad \lambda \geq 4CM_0^3\gamma_2, \quad C\gamma_3 M_0^8\left(\frac{1}{\lambda}+\frac{\lambda}{k^2}\right) \leq \min\left\{\frac{\gamma_2}{8M_0}, \frac{1}{4M_0^2}, \frac{\gamma_1}{8M_0}\right\}. \tag{4.54}$$

From now on, λ and k are fixed. Let

$$\begin{aligned}\mathcal{H}(t) = \int_0^1 \left(\frac{(\varphi_x^I)^2}{2\tilde{v}^2} + \frac{\lambda}{2}(P_k\varphi_x^I)^2\right)dx + \gamma_2\int_0^1\left[R^I\psi^I + \frac{(\varphi^I)^2}{\tilde{v}}\right]dx \\ + \gamma_1\int_0^1\left[\frac{(\psi_x^I)^2}{2\tilde{v}} + \frac{\lambda}{2}(P_k\psi_x^I)^2 + \frac{\zeta_t - q}{\tilde{v}}\psi_x^I\varphi^I\right]dx + \gamma_3\int_0^1\frac{(\psi^I)^2}{\tilde{v}}dx\end{aligned} \tag{4.55}$$

and

$$\mathcal{D}(t) = \frac{1}{4M_0^2}|\varphi_x^I|^2 + \frac{\gamma_2}{4M_0}|\varphi^I|^2 + \frac{\gamma_1}{4M_0}|\psi_t^I|^2 + \frac{\lambda\gamma_3}{4}|\psi_x^I|^2. \tag{4.56}$$

We have, thanks to (4.53) and (4.54),

$$\frac{d}{dt}\mathcal{H}(t) + \mathcal{D}(t) \leq 0. \tag{4.57}$$

It is clear that there exists a constant ν depending only on M_0 such that $\mathcal{H} \leq \frac{1}{\nu}\mathcal{D}$. Therefore

$$\frac{d}{dt}\mathcal{H}(t) + \nu\mathcal{H}(t) \leq 0, \tag{4.58}$$

and $\mathcal{H}(t) \leq \mathcal{H}(0)\mathrm{e}^{-\nu t}$. Thus there exists a constant K_0 depending only on M_0 such that

$$|\varphi_x^I(\cdot,t)|^2 + |\psi_x^I(\cdot,t)|^2 \leq K_0\mathrm{e}^{-\nu t}\big(|\varphi_x^I(\cdot,0)|^2 + |\psi_x^I(\cdot,0)|^2\big). \tag{4.59}$$

(Note that φ^I and ψ^I satisfy the Poincaré inequality.) Finally, since φ^I and ψ^I satisfy linear equations, we can apply (3.34), which is proved below, and the Riesz–Thorin interpolation theorem to obtain (3.35).

***L^2*-continuous dependence for φ^I and ψ^I.** At this point λ and k are fixed. We multiply the equations satisfied by ψ^I, φ^I, and η by ψ^I, φ^I, and η, respectively, and obtain using $|P_k\psi_{xx}^I| \leq k^2|\psi^I|$ that

$$\frac{d}{dt}\big[|\psi^I|^2 + |\varphi^I|^2 + |\eta|^2\big] \leq CM_0^2[|\psi^I|^2 + |\varphi^I|^2] + CM_0^2k^2|\psi^I|^2 + C|\eta|^2, \tag{4.60}$$

from which (3.34) follows immediately.

***H^1*-estimates for φ^{II} and ψ^{II}.** We multiply the equation satisfied by ψ^{II} by $\frac{\psi_t^{II}}{\tilde{v}}$, the equation satisfied by φ^{II} by φ_{xx}^{II}, and the equation satisfied by η by η. Then we apply

$$|\psi_{xx}^{II}|^2 \leq M_0^2\left|\left(\frac{\psi^{II}}{\tilde{v}}\right)_x\right|^2 + CM_0^2|\psi_x^{II}|^2 \tag{4.61}$$

and

$$\left|\left(\frac{\psi^{II}}{\tilde{v}}\right)_x\right| \le |\psi_t^{II}| + M_0|\varphi_x^{II}| + \lambda k^2|\psi^I| + M_0|\varphi^I| + M_0|\varphi^{II}| + M_0|\eta| \tag{4.62}$$

to obtain (3.36). The details are straightforward.

REFERENCES

[1] L. Amerio and G. Prouse, *Abstract Almost Periodic Functions and Functional Equations*, Van Nostrand, New York, 1971.

[2] A. Babin and B. Nicolaenko, *Exponential attractors of reaction-diffusion systems in an unbounded domain*, J. Dynam. Differential Equations, 7 (1995), pp. 567–590.

[3] V. Chepyzhov and M.I. Vishik, *Attractors of non-autonomous dynamical systems and their dimensions*, J. Math. Pures Appl. (9), 73 (1994), pp. 279–333.

[4] A. Eden, C. Foias, B. Nicolaenko, and R. Temam, *Exponential Attractors for Dissipative Evolution Equations*, RAM Res. Appl. Math. 37, John-Wiley, New York, 1994.

[5] A. Eden and J.M. Rakatoson, *Exponential attractors for doubly nonlinear equation*, J. Math. Anal. Appl., 185 (1994), pp. 321–339.

[6] P. Fabrie and A. Miranville, *Exponential attractors for nonautonomous first-order evolution equation*, Discrete Contin. Dynam. Systems, 4 (1998), pp. 225–240.

[7] E. Feireisl, *Global attractors for the Navier–Stokes equations of three-dimensional compressible flow*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 35–39.

[8] C. Foias and E. Olson, *Finite fractal dimension and Hölder-Lipschitz parametrization*, Indiana Univ. Math. J., 45 (1996), pp. 603–616.

[9] C. Galusinski, M. Hind, and A. Miranville, *Exponential attractors for nonautonomous partially dissipative systems*, Differential Integral Equations, 12 (1999), pp. 1–22.

[10] D. Hoff and M. Ziane, *Compact attractors for the Navier–Stokes equations of one dimensional compressible flow*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 239–244.

[11] D. Hoff and M. Ziane, *The global attractor and finite determining nodes for the Navier–Stokes equations of one dimensional compressible flow with singular initial data*, Indiana Univ. Math. J., 49 (2000), pp. 843–889.

[12] O. Ladyzhenskaya, *Finite dimensionality of bounded invariant sets for the Navier–Stokes system and other dissipative systems*, J. Soviet Math., 28 (1985), pp. 714–726.

[13] R. Mañe, *On the dimension of the compact invariant sets of certain nonlinear maps*, in Lecture Notes in Math. 898, Springer-Verlag, New York, 1981, pp. 230–242.

[14] A Miranville, *Exponential attractor for non-autonomous evolution equations*, Appl. Math. Lett., 11 (1998), pp. 19–22.

[15] A Miranville, *Exponential attractor for a class of evolution equations by a decomposition method*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 145–150.

[16] A Miranville, *Exponential attractor for a class of evolution equations by a decomposition method* II. *The non-autonomous case*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 907–912.

[17] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci. 68, Springer-Verlag, New York, 1988.

SIAM J. MATH. ANAL.
Vol. 34, No. 5, pp. 1064–1083

ON ANALYSIS OF STEADY FLOWS OF FLUIDS WITH SHEAR-DEPENDENT VISCOSITY BASED ON THE LIPSCHITZ TRUNCATION METHOD*

JENS FREHSE†, JOSEF MÁLEK‡, AND MARK STEINHAUER§

Abstract. We deal with a system of partial differential equations describing a steady motion of an incompressible fluid with shear-dependent viscosity and present a new global existence result for $p > \frac{2d}{d+2}$. Here p is the coercivity parameter of the nonlinear elliptic operator related to the stress tensor and d is the dimension of the space. Lipschitz test functions, a subtle splitting of the level sets of the maximal functions for the velocity gradients, and a decomposition of the pressure are incorporated to obtain almost everywhere convergence of the velocity gradients.

Key words. incompressible fluid, power-law fluid, shear-dependent viscosity, existence, weak solution, Lipschitz approximation of $W^{1,p}$-functions

AMS subject classifications. 35J55, 35J65, 35J70, 35Q35, 76D99

PII. S0036141002410988

1. Introduction.

Let Ω be an open bounded set in $\mathbb{R}^d$ with boundary $\partial\Omega$. We study the following problem: For given $\mathbf{f} = (f^1, \dots, f^d) : \Omega \to \mathbb{R}^d$ and $\mathbf{T} : \Omega \times \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ we find $\mathbf{v} = (v^1, \dots, v^d) : \Omega \to \mathbb{R}^d$ and $P : \Omega \to \mathbb{R}$ solving

$$-\operatorname{div} \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v})) + \operatorname{div}(\mathbf{v} \otimes \mathbf{v}) + \nabla P = \mathbf{f} \quad \text{in } \Omega, \tag{1.1}$$

$$\operatorname{div} \mathbf{v} = 0 \quad \text{in } \Omega, \tag{1.2}$$

$$\mathbf{v} = \mathbf{0} \quad \text{on } \partial\Omega, \tag{1.3}$$

where $\mathbf{D}(\mathbf{v})$ denotes the symmetric part of the velocity gradient $\nabla\mathbf{v}$, i.e.,

$$\mathbf{D}(\mathbf{v}) \equiv \frac{1}{2}(\nabla\mathbf{v} + \nabla\mathbf{v}^T) \quad \text{with } D_{ij}(\mathbf{v}) = \frac{1}{2}(\partial_i v^j + \partial_j v^i).$$

The main aim of this paper is to present new existence results to (1.1)–(1.3).

Before specifying the main result precisely we fix our notation and give the assumptions on the form of $\mathbf{T}$. The formulation of the main theorem is then completed by comments on earlier results and methods related to the problem (1.1)–(1.3). The last part of this introductory section is devoted to a physical background of the problem and its significance. We also give examples of functions $\mathbf{T}$ and show that they satisfy the assumptions of the main theorem.

*Received by the editors July 10, 2002; accepted for publication (in revised form) September 20, 2002; published electronically April 15, 2003. This work was supported by the SFB256 and SFB611 at the University of Bonn.

http://www.siam.org/journals/sima/34-5/41098.html

†Institute of Applied Mathematics, University of Bonn, Beringstr. 4-6, 53115 Bonn, Germany (erdbeere@iam.uni-bonn.de).

‡Mathematical Institute, Charles University, Sokolovská 83, 18675 Prague 8, Czech Republic (malek@karlin.mff.cuni.cz). This author was also supported by the projects GACR 201/00/0768 and MSM 113200007.

§Mathematical Seminar, University of Bonn, Nussallee 15, 53115 Bonn, Germany (mark@msl.uni-bonn.de).

1.1. Notation. Let $\mathbb{M}$ denote the space of all real $(d \times d)$ matrices $\mathbf{F} = (F_{ij})$, and let $\mathbb{S}$ be its subspace consisting of all symmetric $(d \times d)$ matrices. Using the usual summation convention on repeated indices we set $\mathbf{a} \cdot \mathbf{b} \equiv a^i b^i$ and $(\mathbf{a} \otimes \mathbf{b})_{ij} = a^i b^j$ for $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^d$ and $\mathbf{F} : \mathbf{H} \equiv F_{ij} H_{ij}$ for $\mathbf{F}$, $\mathbf{H} \in \mathbb{M}$. Also we set $|\mathbf{a}| \equiv (\mathbf{a} \cdot \mathbf{a})^{1/2}$ and $|\mathbf{F}| \equiv (\mathbf{F} : \mathbf{F})^{1/2}$.

We use the standard notation of function spaces. If $1 \le q \le +\infty$, then $L^q(\Omega)$ and $W^{k,q}(\Omega)$ ($\mathring{W}^{k,q}(\Omega)$) denote the usual Lebesgue and Sobolev spaces of scalar-, vector-, and tensor-valued functions (with zero traces at the boundary $\partial\Omega$). The norm of $u \in W^{k,q}(\Omega)$ is defined as $\|u\|^q_{k,q;\Omega} \equiv \sum_{|\alpha| \le k} \int_\Omega |D^\alpha u|^q \, dx$.

By $W^{-1,p'}(\Omega)$ we mean the dual space $(\mathring{W}^{1,p}(\Omega))'$ to $\mathring{W}^{1,p}(\Omega)$ with corresponding duality pairing $\langle ., . \rangle_{1,p,\Omega}$.

As usual, $C_0^\infty(\Omega)$ denotes the set of all C^∞-functions with compact support in Ω, while the space $C^\infty_{0,\sigma}(\Omega)$ consists of $\mathbf{\Phi} \in C_0^\infty(\Omega)$ such that $\operatorname{div} \mathbf{\Phi} = 0$. For $p, q \ge 1$ we set

$$
\begin{aligned}
H_q &\equiv \overline{C^\infty_{0,\sigma}(\Omega)}^{\|\cdot\|_{0,q}} = \{\mathbf{v} \in L^q(\Omega) \, : \operatorname{div} \mathbf{v} = 0, \mathbf{v} \cdot \mathbf{n} = 0 \text{ at } \partial\Omega\}, \\
V_p &\equiv \overline{C^\infty_{0,\sigma}(\Omega)}^{\|\nabla\cdot\|_{0,p}} = \{\mathbf{v} \in \mathring{W}^{1,p}(\Omega) \, : \operatorname{div} \mathbf{v} = 0\}, \\
V_p^{'} &\equiv \text{ dual of } V_p.
\end{aligned}
$$

The brackets $\langle ., . \rangle_{V_p}$ represent the duality pairing between V_p and $V_p^{'}$.

If $\mathbf{g}, \mathbf{h}$ are vector-valued functions and $g_i h_i \in L^1(\Omega)$, then $(\mathbf{g}, \mathbf{h}) \equiv \int_\Omega \mathbf{g} \cdot \mathbf{h} \, dx$. Analogously, for tensor-valued functions $\boldsymbol{\eta}, \boldsymbol{\xi}$ satisfying $\eta_{ij} \xi_{ij} \in L^1(\Omega)$ we set $(\boldsymbol{\eta}, \boldsymbol{\xi}) \equiv \int_\Omega \boldsymbol{\eta} : \boldsymbol{\xi} \, dx$.

We will also use the Korn inequality (see [29] for a proof), saying that for $1 < p < +\infty$ there exists a constant $K_p = K_p(\Omega)$ such that

$$
\|\nabla \mathbf{v}\|_{0,p} \le K_p \, \|\mathbf{D}(\mathbf{v})\|_{0,p} \quad \text{for all } \mathbf{v} \in \mathring{W}^{1,p}(\Omega). \tag{1.4}
$$

1.2. Assumptions and main theorem. We start with the formulation of the assumptions on $\mathbf{T} = (T_{ij}) \in \mathbb{S}$.

We assume that $\mathbf{T}$ is a Carathéodory function (i.e., for each fixed $\mathbf{F} \in \mathbb{S}$ the function $x \mapsto \mathbf{T}(x, \mathbf{F})$ is (Lebesgue-) measurable in Ω and the function $\mathbf{F} \mapsto \mathbf{T}(x, \mathbf{F})$ is continuous in $\mathbb{S}$ for almost every $x \in \Omega$) and satisfies for some $p > 1$ the following conditions:

- p-coercivity: there are $c_1 > 0$ and $\varphi_1 \in L^1(\Omega)$ such that

$$
\mathbf{T}(x, \boldsymbol{\eta}) : \boldsymbol{\eta} \ge c_1 |\boldsymbol{\eta}|^p - \varphi_1(x) \tag{1.5}
$$

  for almost all $x \in \Omega$ and for all $\boldsymbol{\eta} \in \mathbb{S}$;
- polynomial growth of order $p - 1$: there are $c_2 > 0$ and $\varphi_2 \in L^{\frac{p}{p-1}}(\Omega)$ such that

$$
|\mathbf{T}(x, \boldsymbol{\eta})| \le c_2 |\boldsymbol{\eta}|^{p-1} + \varphi_2(x) \tag{1.6}
$$

  for almost all $x \in \Omega$ and for all $\boldsymbol{\eta} \in \mathbb{S}$;
- strict monotonicity:

$$
(\mathbf{T}(x, \boldsymbol{\eta}) - \mathbf{T}(x, \boldsymbol{\xi})) : (\boldsymbol{\eta} - \boldsymbol{\xi}) > 0 \tag{1.7}
$$

  for almost all $x \in \Omega$ and for all $\boldsymbol{\eta}, \boldsymbol{\xi} \in \mathbb{S}$ such that $\boldsymbol{\eta} \ne \boldsymbol{\xi}$.

Next, assume that $\mathbf{f} \in W^{-1,p'}(\Omega)$ and (1.5)–(1.7) hold. We say that $\mathbf{v} \in V_p$ is a weak solution to problem (1.1)–(1.3) if

$$\int_\Omega \mathbf{T}(x, \mathbf{D}(\mathbf{v})) : \mathbf{D}(\mathbf{\Phi})\, dx = \langle \mathbf{f}, \mathbf{\Phi} \rangle_{1,p} + \int_\Omega (\mathbf{v} \otimes \mathbf{v}) : \mathbf{D}(\mathbf{\Phi})\, dx \quad \text{for all } \mathbf{\Phi} \in C^\infty_{0,\sigma}(\Omega). \tag{1.8}$$

Note that $(\mathbf{v} \otimes \mathbf{v})_{ij} \equiv v^i v^j \in L^1(\Omega)$ for $p \geq \frac{2d}{d+2}$ due to Sobolev's embedding theorem.

Now we are ready to formulate our existence theorem.

THEOREM 1.1. *Let $p > \frac{2d}{d+2}$, $d \geq 2$. Let $\Omega \subset \mathbb{R}^d$ be an open, bounded set with $\partial\Omega$ of the class $C^{1,1}$. Assume that $\mathbf{f} \in W^{-1,p'}(\Omega)$ and* (1.5)–(1.7) *hold. Then there exists $\mathbf{v} \in V_p$, being a weak solution to* (1.1)–(1.3).

The proof of Theorem 1.1 is split into three parts, each presented in a separate section. In section 2, we introduce suitable approximations to (1.1) and study their basic properties (energy estimates and their consequences, existence of the pressure). Then, in section 3, we present a subtle decomposition of the pressure suitable to our analysis. Finally, in section 4, we present the passage from the solutions of the approximative problems to the solution of (1.1)–(1.3), thus completing the proof.

1.3. Historical comments. Let us first remark that if the convective term $\operatorname{div}(\mathbf{v} \otimes \mathbf{v})$ is neglected in (1.1) and the tensor $\mathbf{T} = (T_{ij})_{i,j=1}^d$ has a potential, i.e., $T_{ij} = \frac{\partial \Phi}{\partial D_{ij}}$, a variational approach can be used. Then the existence of a weak solution can be easily established for all $p > 1$. We refer to the recent works of Fuchs and Seregin [14], [15], where in particular regularity questions for these kinds of problems are discussed.

For proving existence of a weak solution to (1.1)–(1.3) two different methods have been developed. The first one combines the arguments of the standard monotone operator theory with the compactness for $\mathbf{v}$, which turns out to be applicable to (1.1)–(1.3) if $p \geq \frac{3d}{d+2}$. It was performed by Lions [22] and Ladyzhenskaya [18], [19], [20] in the late sixties. The second method, which we call the L^∞-truncation method, yields existence of a weak solution if $p \geq \frac{2d}{d+1}$. It is based on the construction of a special (bounded) test function, a precise characterization of the pressure, and also relies strongly on the strict monotonicity of $\mathbf{T}$. This method was successfully applied to the steady problem in [12] and [32] (in [32] the limiting case $p = \frac{2d}{d+1}$ is not included).

In the present paper we introduce yet another approach, which we call the Lipschitz truncation method, in order to prove the existence of a weak solution for $p > \frac{2d}{d+2}$. We construct a Lipschitz test function to show that for conveniently introduced approximations $\mathbf{v}^n$ we can find a subsequence $\{\mathbf{v}^k\} \subset \{\mathbf{v}^n\}$ such that $\mathbf{D}(\mathbf{v}^k)$ converge almost everywhere to their weak limit $\mathbf{D}(\mathbf{v})$, which is the crucial point in proving that $\mathbf{v}$ is a weak solution to (1.1)–(1.3). *Note that because of earlier results mentioned above we can restrict ourselves within the proof to the case $p \in (\frac{2d}{d+2}, \frac{2d}{d+1}]$.*

Lipschitz truncations of Sobolev functions were already successfully used in different contexts; see [1], [2], [8], [9], [21], [28], [39], [40], [41], and [42]. The novelty of our application of the Lipschitz approximations of Sobolev functions consists of discovering the mechanism of obtaining almost everywhere convergence of gradients for weakly convergent sequences.

Finally we mention that the corresponding time-dependent system is treated in [4], [13], [18], [19], [20], [22], [23], [24], [25], [26], and [30]. We are not going to discuss the dependence of the known existence results on p. We wish to emphasize, however, that we believe that a convenient, probably not straightforward, modification

of the techniques presented in this paper can also improve the existence results for the evolutionary model.

1.4. Continuum mechanical background. Consider isothermal steady flows of an incompressible fluid with a constant density $\rho > 0$. Such flows in a fixed domain Ω are described by the system of equations

$$\operatorname{div} \mathbf{v} = 0, \tag{1.9}$$

$$\rho v^k \frac{\partial \mathbf{v}}{\partial x_k} = \rho \mathbf{f} + \operatorname{div} \mathbf{S}, \tag{1.10}$$

where $\mathbf{v} = (v^1, \dots, v^d)$ is the velocity, $\mathbf{f} = (f^1, \dots, f^d)$ is the density of the volume forces acting on the fluid, and $\mathbf{S}$ is the Cauchy stress.

Equation (1.9) expresses the fact that the fluid is incompressible. Note that due to (1.9) and the fact that the density is constant, the balance of mass is fulfilled.

Equation (1.10) represents the balance of linear momentum. Setting

$$P \equiv -\frac{1}{d} \operatorname{tr} \mathbf{S}, \tag{1.11}$$

we see that the tensor

$$\mathbf{S}_D \equiv \mathbf{S} + P\mathbf{I} \tag{1.12}$$

satisfies

$$\operatorname{tr} \mathbf{S}_D = 0. \tag{1.13}$$

Assuming that $\mathbf{S}_D$ is a tensorial function of the velocity gradient and the fluid is isotropic, the principle of material frame indifference then implies that it happens only through its symmetric part $\mathbf{D}(\mathbf{v})$. Thus,

$$\mathbf{S}_D = \mathbf{T}(\mathbf{D}(\mathbf{v})), \tag{1.14}$$

and one observes that if we put (1.12) with (1.14) into (1.10) and divide the result by ρ, one obtains (1.1). (In fact, the form of $\mathbf{T}$ in (1.1) is more general, as it also permits the dependence of $\mathbf{T}$ on the spatial variable.) Clearly, $v^k \frac{\partial \mathbf{V}}{\partial x_k} = \operatorname{div}(\mathbf{v} \otimes \mathbf{v})$ due to (1.9).

Consider a subclass of (1.14) defined through

$$\mathbf{T}(\mathbf{D}(\mathbf{v})) = \nu(|\mathbf{D}(\mathbf{v})|^2)\mathbf{D}(\mathbf{v}), \tag{1.15}$$

where ν, being a function of the second invariant[1] of the tensor $\mathbf{D}(\mathbf{v})$, is called the generalized viscosity. Materials with the constitutive equation (1.15) are called *fluids with shear-dependent viscosity*. Note that models (1.14)–(1.15) satisfy the requirement (1.13) thanks to (1.9).

Fluids with shear-dependent viscosity represent an important subclass of non-Newtonian fluids, consisting of fluids that have the ability to shear thin (the generalized viscosity decreases as the shear rate in a simple flow increases) or shear thicken (the viscosity increases with the increasing shear rate). In [31], the interested reader

[1]The second invariant II of $\mathbf{D}(\mathbf{v})$ is defined through $II = 1/2[(\operatorname{tr} \mathbf{D}(\mathbf{v}))^2 - \operatorname{tr} \mathbf{D}^2(\mathbf{v})]$. Since in our case $\operatorname{div} \mathbf{v} = \operatorname{tr} \mathbf{D}(\mathbf{v}) = 0$ we conclude that $II = -1/2 \operatorname{tr} \mathbf{D}^2(\mathbf{v}) = -1/2|\mathbf{D}(\mathbf{v})|^2$.

can find a detailed description of these phenomena that were experimentally observed in various areas of engineering such as blood and food rheology, glaciology, geology, and koloid mechanics. We refer to [26] (see also [25]), where a representative list of references to experimental works confirming the presence of nonconstant viscosities in fluids is provided. The power-law fluids, which enjoy significant attention among engineers and physicists, fall into this category; their constitutive equation takes the form (1.15) with

$$\nu(|\mathbf{D}(\mathbf{v})|^2) = \nu_0 |\mathbf{D}(\mathbf{v})|^{p-2}, \tag{1.16}$$

where $p > 1$ is the so-called power-law exponent and $\nu_0 > 0$. Note that if $p = 2$ in (1.16), then the (generalized) viscosity is constant, the dependence of $\mathbf{T}$ on $\mathbf{D}(\mathbf{v})$ in (1.15) is then linear (i.e., the fluid is by definition Newtonian), and the system (1.9)–(1.10) with $\mathbf{S}$ given by (1.12) reduces to the well-known Navier–Stokes equations. For $1 \le p < 2$, the generalized viscosity in (1.16) decreases with increasing $|\mathbf{D}(\mathbf{v})|^2$ and (1.14)–(1.16) then represent shear thinning fluids, while for $p > 2$ (1.14)–(1.16) represent shear thickening fluids. Typical values of the power-law exponent p used in many areas are of the form $\frac{3}{2}$, $\frac{4}{3}$, $\frac{6}{5}$, etc. This recalls a need to have an existence theory for $p \in [1, 2)$, and this also motivates our interest in this direction.

Other examples of models that are widely used in various areas of engineering are given by

$$\mathbf{T}_1(\mathbf{D}(\mathbf{v})) \equiv \nu_0 |\mathbf{D}(\mathbf{v})|^{p-2}\mathbf{D}(\mathbf{v}) + \mu_\infty \mathbf{D}(\mathbf{v}), \tag{1.17}$$

$$\mathbf{T}_2(\mathbf{D}(\mathbf{v})) \equiv \nu_0 (\mu_0 + |\mathbf{D}(\mathbf{v})|^2)^{\frac{p-2}{2}}\mathbf{D}(\mathbf{v}) + \mu_\infty \mathbf{D}(\mathbf{v}), \tag{1.18}$$

$$\mathbf{T}_3(\mathbf{D}(\mathbf{v})) \equiv \mu_\infty \mathbf{D}(\mathbf{v}) + \mu_1 \operatorname{arcsinh}(|\mathbf{D}(\mathbf{v})|)\frac{\mathbf{D}(\mathbf{v})}{|\mathbf{D}(\mathbf{v})|}, \tag{1.19}$$

where μ_0, μ_1, μ_∞, and ν_0 are (at least) nonnegative constants and $p \ge 1$.

Another interesting issue consists of boundary conditions. Here we suppose that the fluid adheres to the boundary, meaning that (no-slip) boundary conditions

$$\mathbf{v} = \mathbf{0} \quad \text{on} \quad \partial\Omega \tag{1.20}$$

are considered.

One could require another type of requirement on the boundary, as for example Navier's, slip, free, or nonhomogeneous Dirichlet boundary conditions, or consider the problem without boundaries in the whole space or in the spatial periodic setting.

We restrict ourselves in what follows to the Dirichlet boundary condition (1.20) for two reasons:

- If one considers another type of boundary condition, the statements of Theorem 1.1 and its proof do not change essentially if one has at one's disposal the basic energy estimates. Thus, the energy estimates are more important than the type of the chosen boundary conditions.
- Boundary conditions (1.20) seem to be reasonable in many applications.

Nevertheless, the reader might find it worthwhile to look at [11] and [17] for further discussion and the treatment of other boundary conditions.

Let us finish this section by showing that the tensors given by formulas (1.17) and (1.18) satisfy the hypotheses (1.5)–(1.7) of Theorem 1.1 if $\nu_0 > 0$ and μ_0, $\mu_\infty \ge 0$. The third example (1.19) satisfies the assumption $p > \frac{2d}{d+2}$ only if $\mu_\infty > 0$.

Example (1.17). Consider $\mathbf{T}_1(\mathbf{F}) = \nu_0|\mathbf{F}|^{p-2}\mathbf{F} + \mu_\infty\mathbf{F}$ with constants $\nu_0 > 0$ and $\mu_\infty \geq 0$ and $\mathbf{F} \in \mathbb{S}$ (corresponding to $\mathbf{D}(\mathbf{v})$). Obviously $\mathbf{T}_1$ is continuous and satisfies both the growth condition

$$|\mathbf{T}(\mathbf{F})| \leq \nu_0|\mathbf{F}|^{p-1} + \mu_\infty|\mathbf{F}|$$

and the coercivity condition

$$\mathbf{T}(\mathbf{F}) : \mathbf{F} = \nu_0|\mathbf{F}|^p + \mu_\infty|\mathbf{F}|^2 \geq \nu_0|\mathbf{F}|^p.$$

For monotonicity we consider two cases.

Case 1. $p \geq 2$. Then we have

$$(\mathbf{T}_1(\mathbf{F}_1) - \mathbf{T}_1(\mathbf{F}_2)) : (\mathbf{F}_1 - \mathbf{F}_2) \geq \nu_0\gamma_0(p,d)|\mathbf{F}_1 - \mathbf{F}_2|^p + \mu_\infty|\mathbf{F}_1 - \mathbf{F}_2|^2,$$

where we use Lemma 4.4 of [7, p. 13]. This inequality shows not only that $\mathbf{T}_1$ is strictly monotone but also that $\mathbf{T}_1$ is uniformly monotone (see [37, p. 500ff., Def. 25.2]).

Case 2. $1 < p < 2$. We are going to verify that

$$(\mathbf{T}_1(\mathbf{F}_1) - \mathbf{T}_1(\mathbf{F}_2)) : (\mathbf{F}_1 - \mathbf{F}_2) \geq \nu_0\gamma_1(p,d)\frac{|\mathbf{F}_1 - \mathbf{F}_2|^2}{(|\mathbf{F}_1| + |\mathbf{F}_2|)^{2-p}} + \mu_\infty|\mathbf{F}_1 - \mathbf{F}_2|^2.$$

To prove it, it is enough to show that

$$(|\mathbf{a}|^{p-2}\mathbf{a} - |\mathbf{b}|^{p-2}\mathbf{b}) \cdot (\mathbf{a} - \mathbf{b}) \geq \gamma_1(p,d)\frac{|\mathbf{a} - \mathbf{b}|^2}{(|\mathbf{a}| + |\mathbf{b}|)^{2-p}} \quad \text{for } \mathbf{a}, \mathbf{b} \in \mathbb{R}^k,$$

which is due to the computations

$$\begin{aligned}
&(|\mathbf{a}|^{p-2}\mathbf{a} - |\mathbf{b}|^{p-2}\mathbf{b}) \cdot (\mathbf{a} - \mathbf{b}) \\
&= \left(\int_0^1 \tfrac{d}{ds}|s\mathbf{a} + (1-s)\mathbf{b}|^{p-2}(s\mathbf{a} + (1-s)\mathbf{b})\, ds\right) \cdot (\mathbf{a} - \mathbf{b}) \\
&= \int_0^1 |s\mathbf{a} + (1-s)\mathbf{b}|^{p-2}|\mathbf{a} - \mathbf{b}|^2\, ds \\
&\qquad + \int_0^1 (p-2)|s\mathbf{a} + (1-s)\mathbf{b}|^{p-2}\left((\mathbf{a} - \mathbf{b}) \cdot \frac{s\mathbf{a} + (1-s)\mathbf{b}}{|s\mathbf{a} + (1-s)\mathbf{b}|}\right)^2 ds \\
&\geq (1 + \min(0, p-2)) \int_0^1 |s\mathbf{a} + (1-s)\mathbf{b}|^{p-2}\, ds\, |\mathbf{a} - \mathbf{b}|^2 \\
&\geq (p-1)\, \frac{|\mathbf{a} - \mathbf{b}|^2}{(|\mathbf{a}| + |\mathbf{b}|)^{2-p}},
\end{aligned}$$

where we use the fact that $1 < p < 2$ and $|s\mathbf{a} + (1-s)\mathbf{b}| \leq |\mathbf{a}| + |\mathbf{b}|$ for all $s \in [0,1]$. Thus, we conclude that $\mathbf{T}_1$ is strictly monotone but in general not uniformly monotone.

Example (1.18). Consider $\mathbf{T}_2(\mathbf{F}) = \nu_0(\mu_0 + |\mathbf{F}|^2)^{\frac{p-2}{2}}\mathbf{F} + \mu_\infty\mathbf{F}$ with constants $\nu_0 > 0$ and $\mu_0, \mu_\infty \geq 0$. Computations similar to the previous example show that all hypotheses of Theorem 1.1 are satisfied (see [25, Chap. 5, pp. 193–196, 198ff., Lem. 1.19]).

2. Approximations and their properties. We define approximations to our problem in the following way: For $m = 1, 2, 3, \dots, p > 1$ and $q \geq \frac{2p}{p-1} = 2p'$ we look for $(\mathbf{v}^m, P^m)$ solving in Ω

$$\begin{aligned} -\operatorname{div} \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}^m)) + \operatorname{div}(\mathbf{v}^m \otimes \mathbf{v}^m) + \frac{1}{m}|\mathbf{v}^m|^{q-2}\mathbf{v}^m &= \mathbf{f} - \nabla P^m, \\ \operatorname{div} \mathbf{v}^m &= 0, \end{aligned} \tag{2.1}$$

complemented by the boundary conditions

$$\mathbf{v}^m = \mathbf{0} \quad \text{on } \partial\Omega. \tag{2.2}$$

The following lemma can be proved by standard arguments of the monotone operator theory and via the compact embedding $\mathring{W}^{1,p}(\Omega) \hookrightarrow\hookrightarrow L^2(\Omega)$ valid for $p > \frac{2d}{d+2}$.

LEMMA 2.1. *Let $p > \frac{2d}{d+2}$ and $q \geq \frac{2p}{p-1} = 2p'$. Suppose $\mathbf{f} \in W^{-1,p'}(\Omega)$. Then there exists $\mathbf{v}^m \in V_p \cap H_q$ satisfying*

$$\begin{aligned} \int_\Omega \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}^m)) : \mathbf{D}(\mathbf{\Phi})\, dx + \frac{1}{m}\int_\Omega |\mathbf{v}^m|^{q-2}\mathbf{v}^m \cdot \mathbf{\Phi}\, dx &= \langle \mathbf{f}, \mathbf{\Phi} \rangle_{1,p} \\ + \int_\Omega (\mathbf{v}^m \otimes \mathbf{v}^m) : \mathbf{D}(\mathbf{\Phi})\, dx \quad &\text{for all } \mathbf{\Phi} \in C^\infty_{0,\sigma}(\Omega). \end{aligned} \tag{2.3}$$

Moreover, all $\mathbf{v}^m$ satisfy the uniform estimate

$$\|\mathbf{D}(\mathbf{v}^m)\|^p_{0,p} + \|\nabla \mathbf{v}^m\|^p_{0,p} + \frac{1}{m}\|\mathbf{v}^m\|^q_{0,q} \leq K \tag{2.4}$$

and consequently, due to the growth condition (1.6) *and Sobolev's embedding theorem (considered in the interesting case $p < d$),*

$$\|\mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}^m))\|_{0,p'} \leq K, \tag{2.5}$$

$$\|\mathbf{v}^m\|_{0,\frac{dp}{d-p}} \leq K, \tag{2.6}$$

$$\|\mathbf{v}^m \otimes \mathbf{v}^m\|_{0,\frac{dp}{2(d-p)}} \leq K. \tag{2.7}$$

Let us emphasize that due to earlier existence results mentioned in the introduction, it is enough to deal with the case

$$p \in \Big(\frac{2d}{d+2}, \frac{2d}{d+1}\Big].$$

Next, we introduce the (approximative) pressures P^m, observing that in (2.3) we can use test functions $\mathbf{\Phi}$ from $V_p \cap V_r = V_r$, where $\frac{1}{r} = 1 + \frac{2}{d} - \frac{2}{p} = \frac{(d+2)p-2d}{dp}$ because of $V_p \hookrightarrow L^{\frac{dp}{d-p}}$. Let us note that for all $s \in [1, \infty)$, $V_r \hookrightarrow\hookrightarrow L^s$ for $\frac{2d}{d+2} < p \leq \frac{2d}{d+1}$. Defining the functional $\mathbf{F}^m$ as

$$\begin{aligned} \langle \mathbf{F}^m, \mathbf{\Phi} \rangle_{1,r,\Omega} \equiv \int_\Omega \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}^m)) : \mathbf{D}(\mathbf{\Phi})\, dx + \frac{1}{m}\int_\Omega |\mathbf{v}^m|^{q-2}\mathbf{v}^m \cdot \mathbf{\Phi}\, dx \\ - \int_\Omega (\mathbf{v}^m \otimes \mathbf{v}^m) : \mathbf{D}(\mathbf{\Phi})\, dx - \langle \mathbf{f}, \mathbf{\Phi} \rangle_{1,p,\Omega} \end{aligned} \tag{2.8}$$

we see that $\langle \mathbf{F}^m, \mathbf{\Phi}\rangle_{1,r,\Omega} = 0$ for all $\mathbf{\Phi} \in C^\infty_{0,\sigma}(\Omega)$ due to (2.3). Moreover $\mathbf{F}^m \in W^{-1,r'}(\Omega)$ and

$$\|\mathbf{F}^m\|_{-1,r'} \le K \quad \text{with } r' = \frac{r}{r-1} = \frac{dp}{dp-(d+2)p+2d} = \frac{dp}{2(d-p)}\,.$$

By a version of De Rham's theorem (see, for example, [3, Thm. 2.8, p. 116ff.]) there exists $P^m \in L^{r'}(\Omega)$ with zero mean value over each connected component of Ω such that

$$\langle \mathbf{F}^m, \mathbf{\Phi}\rangle_{1,r,\Omega} \equiv \langle -\nabla P^m, \mathbf{\Phi}\rangle_{1,r,\Omega} = \int_\Omega P^m \operatorname{div} \mathbf{\Phi}\, dx \tag{2.9}$$

and

$$\|P^m\|_{0,r'} \le C\, \|\nabla P^m\|_{-1,r'} \le C\, \|F^m\|_{-1,r'} \le K. \tag{2.10}$$

As a consequence of these observations we obtain the equivalent weak formulation to (2.3),

$$\begin{aligned}\int_\Omega \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}^m)) : \mathbf{D}(\mathbf{\Phi})\, dx + \frac{1}{m}\int_\Omega |\mathbf{v}^m|^{q-2}\mathbf{v}^m \cdot \mathbf{\Phi}\, dx \\ = \langle \mathbf{f}, \mathbf{\Phi}\rangle_{1,p} + \int_\Omega (\mathbf{v}^m \otimes \mathbf{v}^m) : \mathbf{D}(\mathbf{\Phi})\, dx + \int_\Omega P^m \operatorname{div} \mathbf{\Phi}\, dx,\end{aligned} \tag{2.11}$$

valid for all $m = 1, 2, 3, \dots$ and all $\mathbf{\Phi} \in \mathring{W}^{1,r}(\Omega)$ with $r = \frac{dp}{(d+2)p-2d}$. Note that if $p \in (\frac{2d}{d+2}, \frac{2d}{d+1}]$, then $r \ge d$.

The uniform estimates (2.4)–(2.7) and (2.10) imply the existence of a subsequence $\{(\mathbf{v}^k, P^k)\}_{k\in\mathbb{N}} = \{(\mathbf{v}^{m_k}, P^{m_k})\}_{k\in\mathbb{N}}$ of $\{(\mathbf{v}^m, P^m)\}_{m\in\mathbb{N}}$ and $(\mathbf{v}, P) \in V_p \times L^{r'}(\Omega)$ such that $(k \to \infty)$

$$\mathbf{D}(\mathbf{v}^k) \rightharpoonup \mathbf{D}(\mathbf{v}) \qquad \text{weakly in } L^p(\Omega), \tag{2.12}$$

$$\nabla \mathbf{v}^k \rightharpoonup \nabla \mathbf{v} \qquad \text{weakly in } L^p(\Omega), \tag{2.13}$$

$$\mathbf{v}^k \to \mathbf{v} \qquad \text{strongly in } L^s(\Omega) \text{ for all } s \in [1, 2r'), \tag{2.14}$$

$$\mathbf{v}^k \to \mathbf{v} \qquad \text{almost everywhere in } \Omega, \tag{2.15}$$

$$\mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}^k)) \rightharpoonup \boldsymbol{\chi} \qquad \text{weakly in } L^{p'}(\Omega), \tag{2.16}$$

$$P^k \rightharpoonup P \qquad \text{weakly in } L^{r'}(\Omega). \tag{2.17}$$

Now we want to pass to the limit in (2.11) as $k \to \infty$. In order to do so we first observe that (2.4) implies for every $\mathbf{\Phi} \in C^\infty_0(\Omega)$ and $k \to \infty$

$$\left|\frac{1}{k}\int_\Omega |\mathbf{v}^k|^{q-2}\mathbf{v}^k \cdot \mathbf{\Phi}\, dx\right| \le \frac{1}{k^{1/q}}\left(\frac{1}{k}\|\mathbf{v}^k\|_q^q\right)^{\frac{q-1}{q}} \|\mathbf{\Phi}\|_q \ \to 0. \tag{2.18}$$

The convective term is treated with help from the compact embedding $\mathring{W}^{1,p} \hookrightarrow\hookrightarrow L^2$, $p > \frac{2d}{d+2}$. Writing $\mathbf{v}^k = \mathbf{v} + \mathbf{v}^k - \mathbf{v}$ we have for every $\mathbf{\Phi} \in C^\infty_0(\Omega)$ and $k \to \infty$

$$\begin{aligned}\int_\Omega (\mathbf{v}^k \otimes \mathbf{v}^k) : \mathbf{D}(\mathbf{\Phi})\, dx &= \int_\Omega [(\mathbf{v}^k - \mathbf{v}) \otimes \mathbf{v}^k] : \mathbf{D}(\mathbf{\Phi})\, dx \\ &\quad + \int_\Omega [\mathbf{v} \otimes (\mathbf{v}^k - \mathbf{v})] : \mathbf{D}(\mathbf{\Phi})\, dx + \int_\Omega (\mathbf{v} \otimes \mathbf{v}) : \mathbf{D}(\mathbf{\Phi})\, dx \\ &\to \int_\Omega (\mathbf{v} \otimes \mathbf{v}) : \mathbf{D}(\mathbf{\Phi})\, dx.\end{aligned} \tag{2.19}$$

Owing to (2.17) we also observe that for $\boldsymbol{\Phi} \in C_0^\infty(\Omega)$ and $k \to \infty$

$$\int_\Omega P^k \mathrm{div}\, \boldsymbol{\Phi}\, dx \quad \to \quad \int_\Omega P \mathrm{div}\, \boldsymbol{\Phi}\, dx. \tag{2.20}$$

Collecting our results we find that $\mathbf{v} \in V_p$ satisfies

$$\int_\Omega \boldsymbol{\chi} : \mathbf{D}(\boldsymbol{\Phi})\, dx = \langle \mathbf{f}, \boldsymbol{\Phi} \rangle_{1,p} + \int_\Omega (\mathbf{v} \otimes \mathbf{v}) : \mathbf{D}(\boldsymbol{\Phi})\, dx + \int_\Omega P \mathrm{div}\, \boldsymbol{\Phi}\, dx \tag{2.21}$$

for all $\boldsymbol{\Phi} \in C_0^\infty(\Omega)$, respectively, $\boldsymbol{\Phi} \in \mathring{W}^{1,r}(\Omega)$.

Our aim now is to demonstrate that $\boldsymbol{\chi} = \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}))$. For this purpose it suffices to show that

$$\mathbf{D}(\mathbf{v}^k) \to \mathbf{D}(\mathbf{v}) \qquad \text{in measure on } \Omega$$

or almost everywhere convergence on compact subsets of Ω. If this were true, then we could find a further subsequence by a diagonal procedure (for simplicity we do not change notation) such that

$$\mathbf{D}(\mathbf{v}^k) \to \mathbf{D}(\mathbf{v}) \qquad \text{almost everywhere in } \Omega. \tag{2.22}$$

Then, by Vitali's theorem (with help from the growth condition (1.6)) we obtain

$$\int_\Omega \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}^k)) : \mathbf{D}(\boldsymbol{\Phi})\, dx \quad \to \quad \int_\Omega \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v})) : \mathbf{D}(\boldsymbol{\Phi})\, dx \tag{2.23}$$

and we can finish the proof of Theorem 1.1.

Note also that once we have (2.22) then we easily conclude from (2.4), respectively, (2.12), using Vitali's theorem that

$$\mathbf{D}(\mathbf{v}^k) \to \mathbf{D}(\mathbf{v}) \qquad \text{strongly in } L^s(\Omega) \qquad \text{for all } s \in [1, p),$$

which is due to (1.4) tantamount to

$$\mathbf{v}^k \to \mathbf{v} \qquad \text{strongly in } \mathring{W}^{1,s}(\Omega) \qquad \text{for all } s \in [1, p).$$

The missing proof of (2.22) will be given in section 4, while the next section is devoted to a decomposition of the pressure P^k.

3. Decomposition of the pressure. Consider four auxiliary Stokes problems, $I = 1, 2, 3, 4$,

$$\begin{aligned} -\Delta \mathbf{u}^{I_k} + \nabla P^{I_k} &= \mathbf{H}^{I_k} && \text{in } \Omega, \\ \mathrm{div}\, \mathbf{u}^{I_k} &= 0 && \text{in } \Omega, \\ \mathbf{u}^{I_k} &= \mathbf{0} && \text{on } \partial\Omega, \end{aligned} \tag{3.1}$$

where

$$\begin{aligned} \mathbf{H}^{1_k} &= -\mathrm{div}\, \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}^k)) \in (\mathring{W}^{1,p}(\Omega))^*, \\ \mathbf{H}^{2_k} &= \mathrm{div}\, (\mathbf{v}^k \otimes (\mathbf{v}^k - \mathbf{v})) \in (\mathring{W}^{1,r}(\Omega))^*, \\ \mathbf{H}^{3_k} &= \mathrm{div}\, ((\mathbf{v}^k - \mathbf{v}) \otimes \mathbf{v}) \in (\mathring{W}^{1,r}(\Omega))^*, \\ \mathbf{H}^{4_k} &= \frac{1}{k} |\mathbf{v}^k|^{q-2} \mathbf{v}^k \in (L^q(\Omega))^*. \end{aligned} \tag{3.2}$$

The classical theory for the Stokes system (cf. [3], for example) implies the existence of solutions $(\mathbf{u}^{I_k}, P^{I_k})$, $I = 1,2,3,4$, with the following estimates on the pressures P^{I_k} having zero mean value over each connected component of Ω:

$$\|P^{1_k}\|_{0,p';\Omega} \le C\|\mathbf{H}^{1_k}\|_{(\mathring{W}^{1,p}(\Omega))^*} \le C\|\mathbf{T}(\cdot,\mathbf{D}(\mathbf{v}^k))\|_{0,p';\Omega}, \tag{3.3}$$

$$\begin{aligned}\|P^{2_k}\|_{0,r';\Omega} &\le C\|\mathbf{H}^{2_k}\|_{(\mathring{W}^{1,r}(\Omega))^*} \le C\|\mathbf{v}^k\otimes(\mathbf{v}^k-\mathbf{v})\|_{0,r';\Omega}\\ &\le C\|\mathbf{v}^k\|_{0,2r';\Omega}\|\mathbf{v}^k-\mathbf{v}\|_{0,2r';\Omega},\end{aligned} \tag{3.4}$$

$$\|P^{3_k}\|_{0,r';\Omega} \le C\|\mathbf{H}^{3_k}\|_{(\mathring{W}^{1,r}(\Omega))^*} \le C\|\mathbf{v}^k-\mathbf{v}\|_{0,2r';\Omega}\|\mathbf{v}^k\|_{0,2r';\Omega}, \tag{3.5}$$

$$\begin{aligned}\|\nabla P^{4_k}\|_{0,q';\Omega} &\le C\|\mathbf{H}^{4_k}\|_{(L^q(\Omega))^*} \le \frac{C}{k}\||\mathbf{v}^k|^{q-1}\|_{0,q';\Omega}\\ &\le \frac{C}{k^{1/q}}\left(\frac{1}{k^{1/q}}\|\mathbf{v}^k\|_{0,q;\Omega}\right)^{q-1}.\end{aligned} \tag{3.6}$$

As $2r' = \frac{dp}{d-p}$, it follows from (3.4), (3.5), and (2.14) that for $k\to\infty$ we have

$$P^{2_k}\to 0 \text{ and } P^{3_k}\to 0 \quad \text{strongly in } L^s(\Omega) \quad \text{for all } s\in[1,r'). \tag{3.7}$$

Also, due to (2.4) we observe that for $k\to\infty$

$$\nabla P^{4_k}\to 0 \quad \text{strongly in } L^{q'}(\Omega). \tag{3.8}$$

Of course, one has analogous estimates for $\mathbf{u}^{I_k}$. For our purpose, it is enough to know that

$$\mathbf{U}^k \equiv \mathbf{u}^{1_k}+\mathbf{u}^{2_k}+\mathbf{u}^{3_k}+\mathbf{u}^{4_k} \quad \text{with} \quad \operatorname{div}\mathbf{U}^k = 0$$

satisfy

$$\|\mathbf{U}^k\|_{1,r';\Omega} \le K. \tag{3.9}$$

Next, summing up the weak forms of the problems $(3.1)_I$ over $I = 1,2,3,4$ and using (2.11) we obtain

$$\begin{aligned}\int_\Omega \nabla\mathbf{U}^k : \nabla\boldsymbol{\Phi}\,dx - \sum_{I=1}^4 \int P^{I_k}\operatorname{div}\boldsymbol{\Phi}\,dx &= \langle\mathbf{f},\boldsymbol{\Phi}\rangle_{1,p} + \int P^k\operatorname{div}\boldsymbol{\Phi}\,dx\\ &\quad + \int_\Omega(\mathbf{v}\otimes\mathbf{v}) : \mathbf{D}(\boldsymbol{\Phi})\,dx \qquad \text{for all } \boldsymbol{\Phi}\in\mathring{W}^{1,r}(\Omega).\end{aligned} \tag{3.10}$$

Taking $\boldsymbol{\Phi}$ from V_r in (3.10) (i.e., $\operatorname{div}\boldsymbol{\Phi} = 0$) we conclude that

$$\int_\Omega \nabla\mathbf{U}^k : \nabla\boldsymbol{\Phi}\,dx = \langle\mathbf{f},\boldsymbol{\Phi}\rangle_{1,p} + \int_\Omega(\mathbf{v}\otimes\mathbf{v}) : \mathbf{D}(\boldsymbol{\Phi})\,dx \qquad \text{for all } \boldsymbol{\Phi}\in V_r. \tag{3.11}$$

This and (3.9) then imply

$$\mathbf{U}^k = \mathbf{U}\in\mathring{W}^{1,r'}(\Omega) \qquad \text{for all } k\in\mathbb{N}. \tag{3.12}$$

Indeed, it follows from (3.11) that for $k,\ell\in\mathbb{N}$

$$\int_\Omega \nabla(\mathbf{U}^k-\mathbf{U}^\ell) : \nabla\boldsymbol{\Phi}\,dx = 0 \qquad \text{for all } \boldsymbol{\Phi}\in V_r.$$

Choosing $\mathbf{\Phi}$ to be the solution of

$$\begin{aligned} -\Delta\mathbf{\Phi} + \nabla Q &= \frac{\mathbf{U}^k - \mathbf{U}^\ell}{|\mathbf{U}^k - \mathbf{U}^\ell|} && \text{in } \Omega, \\ \operatorname{div}\mathbf{\Phi} &= 0 && \text{in } \Omega, \\ \mathbf{\Phi} &= \mathbf{0} && \text{on } \partial\Omega \end{aligned}$$

leads to (3.12).

Finally, taking (2.11) into account again and replacing $\int_\Omega P^k \operatorname{div}\mathbf{\Phi}\,dx$ with help from (3.10) and (3.12) we obtain

$$\begin{aligned} &\int_\Omega \mathbf{T}(\cdot,\mathbf{D}(\mathbf{v}^k)) : \mathbf{D}(\mathbf{\Phi})\,dx + \frac{1}{k}\int_\Omega |\mathbf{v}^k|^{q-2}\mathbf{v}^k \cdot \mathbf{\Phi}\,dx \\ &\quad = \int_\Omega (\mathbf{v}^k \otimes \mathbf{v}^k) : \mathbf{D}(\mathbf{\Phi})\,dx - \sum_{I=1}^{4}\int_\Omega P^{I_k}\operatorname{div}\mathbf{\Phi}\,dx \\ &\qquad + \int_\Omega \nabla\mathbf{U} : \nabla\mathbf{\Phi}\,dx - \int_\Omega (\mathbf{v}\otimes\mathbf{v}) : \mathbf{D}(\mathbf{\Phi})\,dx \quad \text{for all } \mathbf{\Phi} \in \mathring{W}^{1,r}(\Omega). \end{aligned} \tag{3.13}$$

The advantage of this formulation stems from more precise control of the particular pressures P^{1_k}, P^{2_k}, P^{3_k}, and P^{4_k} owing to (3.3)–(3.8).

4. Almost everywhere convergence of $\mathbf{D}(\mathbf{v}^k)$ to $\mathbf{D}(\mathbf{v})$. The desired convergence of $\mathbf{D}(\mathbf{v}^k)$ to $\mathbf{D}(\mathbf{v})$ almost everywhere in Ω will certainly hold if one shows that for a given but arbitrary $\eta > 0$ there is a subsequence $\{\mathbf{v}^\ell\}_{\ell\in\mathbb{N}} \subset \{\mathbf{v}^k\}_{k\in\mathbb{N}}$ such that (for some $\theta \in (0,1)$, say, $\theta = \frac{1}{2}$)

$$\lim_{\ell\to\infty}\int_\Omega \Big[(\mathbf{T}(\cdot,\mathbf{D}(\mathbf{v}^\ell)) - \mathbf{T}(\cdot,\mathbf{D}(\mathbf{v}))) : \mathbf{D}(\mathbf{v}^\ell - \mathbf{v})\Big]^\theta\,dx \le \eta. \tag{4.1}$$

To reach this goal it seems natural to consider

$$\mathbf{v}^k - \mathbf{v} \tag{4.2}$$

as a test function in (2.3), rewrite the left-hand side of the obtained equality as in (4.1) with $\theta = 1$, and show that the remaining terms are small as $k \to \infty$. Unfortunately, this idea works only for $p \ge \frac{3d}{d+2}$.

In [12], the L^∞-truncation of (4.2), namely,

$$(\mathbf{v}^k - \mathbf{v})\left(1 - \min\left(\frac{|\mathbf{v}^k - \mathbf{v}|}{L}, 1\right)\right) \quad \text{with} \quad L > 0 \quad \text{small}, \tag{4.3}$$

has been successfully applied to deduce (4.1). The main difficulty is showing the smallness of the integral

$$\int_{\Omega_L^k} \mathbf{T}(\cdot,\mathbf{D}(\mathbf{v}^k)) : \mathbf{D}\left((\mathbf{v}^k - \mathbf{v})\left(1 - \min\left(\frac{|\mathbf{v}^k - \mathbf{v}|}{L}, 1\right)\right)\right)\,dx,$$

where $\Omega_L^k \equiv \{x \in \Omega; |\mathbf{v}^k(x) - \mathbf{v}(x)| < L\}$. The L^∞-truncation method works for $p \ge \frac{2d}{d+1}$; the bound is due to the required L^1-integrability of the convective term $(\mathbf{v}\cdot\nabla)\mathbf{v}$.

Following the goal to prove Theorem 1.1 we have observed that it is enough to restrict ourselves to the case $p \in (\frac{2d}{d+2}, \frac{2d}{d+1}]$, and it is necessary to use a smoother test function than in (4.3) in order to control the convective term; yet the test function should not differ from (4.2) too much.

For this purpose, we test (3.13) by

$$(\mathbf{v}^k - \mathbf{v})_\lambda \quad \text{with } \lambda > 0 \quad \text{large enough,} \tag{4.4}$$

using the notation z^k_λ to denote such a Lipschitz (i.e., $\mathring{W}^{1,\infty}$-) truncation of z^k so that z^k_λ coincides with z^k except for a small set A^k_λ.

Let us remark that the idea to approximate a $\mathring{W}^{1,p}$-function $\mathbf{w}$ by a Lipschitz-continuous function which agrees with $\mathbf{w}$ on a "large" set was developed earlier; see [2], [10], [16], [21], [8], and [9], among others.

The proof of (4.1), and consequently of Theorem 1.1, is split into three steps. First, in Proposition 4.1, we study properties of $(\mathbf{v}^k - \mathbf{v})_\lambda$ for general λ. Then we cover the exceptional sets of noncoincidence A^k_λ by two sets F^k_λ and G^k_λ and show (see Propositions 4.1 and 4.3) by fixing λ and taking a convenient subsequence $\{\mathbf{v}^\ell\}_{\ell\in\mathbb{N}}$ that certain quantities are small on these sets. Finally, we prove (4.1) in Proposition 4.4.

PROPOSITION 4.1. *There is a constant $C = C(\Omega, d)$ such that whenever $\mathbf{w}^m \rightharpoonup 0$ weakly in $\mathring{W}^{1,p}(\Omega)$, then for all $\lambda > 0$ there is a sequence $\{\mathbf{w}^m_\lambda\}_{m\in\mathbb{N}} \subset \mathring{W}^{1,\infty}(\Omega)$ such that*

$$\|\mathbf{w}^m_\lambda\|_{1,\infty;\,\Omega} \le C\lambda. \tag{4.5}$$

Moreover, denoting $A^m_\lambda \equiv \{x \in \Omega; \mathbf{w}^m_\lambda(x) \neq \mathbf{w}^m(x)\}$, we then obtain

$$|A^m_\lambda| \le \frac{C}{\lambda^p}\|\nabla \mathbf{w}^m\|^p_{0,p;\,\Omega}. \tag{4.6}$$

Consequently,

$$\|\nabla \mathbf{w}^m_\lambda\|^p_{0,p;\,\Omega} \le C\|\nabla \mathbf{w}^m\|^p_{0,p;\,\Omega} \le K \tag{4.7}$$

and (as $m \to \infty$)

$$\begin{aligned} &\mathbf{w}^m_\lambda \to 0 \quad \textit{strongly in } L^s(\Omega) \qquad \textit{for all } s \in [1,\infty), \\ &\mathbf{w}^m_\lambda \rightharpoonup 0 \quad \textit{weakly in } \mathring{W}^{1,s}(\Omega) \qquad \textit{for all } s \in [1,\infty). \end{aligned} \tag{4.8}$$

In addition, we construct sets F^m_λ and G^m_λ such that

$$|A^m_\lambda| \le |F^m_\lambda| + |G^m_\lambda|, \tag{4.9}$$

$$|F^m_\lambda| \le \frac{C}{\lambda^p}\|\nabla \mathbf{w}^m\|^p_{0,p;\,\Omega}, \tag{4.10}$$

$$|G^m_\lambda| \le \frac{C}{\lambda^{2p}}\|\nabla \mathbf{w}^m\|^p_{0,p;\,\Omega}. \tag{4.11}$$

Before providing a proof of this proposition we recall Kirszbraun's extension theorem (see [21, Prop. 2.1, p. 708]).

LEMMA 4.2. *Let $\mathcal{M}$ be a metric space and let $\mathcal{K}$ be a subset of $\mathcal{M}$ such that $u : \mathcal{K} \to \mathbb{R}$ is Lipschitz-continuous with Lipschitz constant L. Then there exists a*

continuation $\hat{u} : \mathcal{M} \to \mathbb{R}$ of u such that $\hat{u}$ is Lipschitz-continuous with the same Lipschitz bound L and

$$\sup_{x\in\mathcal{M}} |\hat{u}(x)| \le \sup_{x\in\mathcal{K}} |u(x)|.$$

Proof of Proposition 4.1. The proof is based on ideas from [21, Prop. 2.2, p. 709] and [9, Lem. 4.1, pp. 21–22]. Extending $\mathbf{w}^m$ by zero we obtain $\tilde{\mathbf{w}}^m \in \tilde{W}^{1,p}(\mathbb{R}^d) = W^{1,p}(\mathbb{R}^d)$ with $\tilde{\mathbf{w}}^m \rightharpoonup 0$ weakly in $W^{1,p} = W^{1,p}(\mathbb{R}^d)$.

Recalling the definition of the Hardy–Littlewood maximal function of $\nabla\tilde{\mathbf{w}}^m$,

$$M(\nabla\tilde{\mathbf{w}}^m)(x) \equiv \sup_{r>0} \frac{1}{|B_r(x)|} \int_{B_r(x)} |\nabla\tilde{\mathbf{w}}^m(y)|\, dy,$$

we define for $\lambda > 1$

$$R^m_\lambda \equiv F^m_\lambda \cup G^m_\lambda \cup \left\{x \in \mathbb{R}^d : x \text{ is not a Lebesgue point of } \nabla\tilde{\mathbf{w}}^m\right\}, \tag{4.12}$$

where

$$\begin{aligned} F^m_\lambda &\equiv \left\{x \in \mathbb{R}^d : \lambda < M(\nabla\tilde{\mathbf{w}}^m)(x) \le \lambda^2\right\}, \\ G^m_\lambda &\equiv \left\{x \in \mathbb{R}^d : M(\nabla\tilde{\mathbf{w}}^m)(x) > \lambda^2\right\}. \end{aligned} \tag{4.13}$$

Note that the Lebesgue measure of the last set in the definition of R^m_λ is zero.

Since $M : L^p \to L^p$ is a "bounded" operator (see, for example, [34, pp. 4–12] or [38, Thm. 2.8.2, p. 84ff.]) we obtain

$$\begin{aligned} \lambda|R^m_\lambda| &\le \int_{R^m_\lambda} M(\nabla\tilde{\mathbf{w}}^m(x))\, dx \le \|M(\nabla\tilde{\mathbf{w}}^m)\|_{0,p} |R^m_\lambda|^{1-\frac{1}{p}} \\ &\le C\|\nabla\tilde{\mathbf{w}}^m\|_{0,p} |R^m_\lambda|^{1-\frac{1}{p}}, \end{aligned}$$

which implies

$$|R^m_\lambda| \le \frac{C}{\lambda^p} \|\nabla\tilde{\mathbf{w}}^m\|^p_{0,p} \quad \text{and} \quad |F^m_\lambda| \le \frac{C}{\lambda^p} \|\nabla\tilde{\mathbf{w}}^m\|^p_{0,p}. \tag{4.14}$$

Analogously, one obtains

$$|G^m_\lambda| \le \frac{C}{\lambda^{2p}} \|\nabla\tilde{\mathbf{w}}^m\|^p_{0,p}. \tag{4.15}$$

Next, from Lemma 1 in [2] it follows that there is a constant $C(d)$ such that

$$|\tilde{\mathbf{w}}^m(x) - \tilde{\mathbf{w}}^m(y)| \le C(d)\, \lambda\, |x - y| \quad \text{on } \mathbb{R}^d \setminus R^\lambda_m \tag{4.16}$$

and

$$|\tilde{\mathbf{w}}^m(x) - (\tilde{\mathbf{w}}^m)_{x,r}| \le C(d)\, r\, \lambda \quad \text{on } \mathbb{R}^d \setminus R^\lambda_m,$$

where $(\tilde{\mathbf{w}}^m)_{x,r}$ is the mean value of $\tilde{\mathbf{w}}^m$ over $B_r(x)$. Choosing $x \in \Omega \setminus R^\lambda_m$ and $r = 2\,\mathrm{dist}(x, \Omega^C)$, the smoothness of the boundary[2] implies the existence of A (independent of x) such that

$$|B_r(x) \cap \Omega^C| \ge A\, r^d.$$

[2] In fact, it would be sufficient for this part of the proof to assume that the boundary of Ω satisfies the cone property.

Hence Poincaré's inequality yields

$$|(\tilde{\mathbf{w}}^m)_{x,r}| \le C\, r \sup_{r>0} \frac{1}{|B_r(x)|} \int_{B_r(x)} |\nabla \tilde{\mathbf{w}}^m(y)|\, dy \le C \operatorname{dist}(x, \Omega^C)\, \lambda.$$

Thus

$$|\tilde{\mathbf{w}}^m(x)| \le C \operatorname{dist}(x, \Omega^C)\, \lambda \quad \text{on } \mathbb{R}^d \setminus R^\lambda_m.$$

This implies that

$$\tilde{\mathbf{w}}^m_\lambda(x) \equiv \begin{cases} \mathbf{w}^m(x) & \text{on } \Omega \setminus (R^\lambda_m), \\ 0 & \text{on } \mathbb{R}^d \setminus \Omega \end{cases}$$

is bounded and Lipschitz-continuous on its domain of definition. Thus by Lemma 4.2 there exists an extension $\mathbf{w}^m_\lambda$ to $\mathbb{R}^d$ with Lipschitz constant $C(d)\,\lambda$ and L^∞-bound $C\rho\lambda$, where ρ denotes the diameter of Ω. The assertion (4.5) is proved.

Moreover, the set $A^m_\lambda = \{x \in \Omega; \mathbf{w}^m_\lambda(x) \ne \mathbf{w}^m(x)\}$ is a subset of R^m_λ and $|A^m_\lambda| \le |R^m_\lambda|$. This together with (4.12)–(4.15) yields (4.6) and (4.9)–(4.11). Further, by (4.14) we have

$$\begin{aligned} \|\nabla \mathbf{w}^m_\lambda\|_{0,p;\,\Omega} &= \|\nabla \mathbf{w}^m_\lambda\|_{0,p;\,\Omega\setminus R^m_\lambda} + \|\nabla \mathbf{w}^m_\lambda\|_{0,p;\,R^m_\lambda} \\ &\le \|\nabla \mathbf{w}^m\|_{0,p;\,\Omega\setminus R^m_\lambda} + C\lambda |R^m_\lambda|^{1/p} \\ &\le \|\nabla \mathbf{w}^m\|_{0,p;\,\Omega\setminus R^m_\lambda} + C\|\nabla \mathbf{w}^m\|_{0,p;\,\Omega} \le (C+1)\|\nabla \mathbf{w}^m\|_{0,p;\,\Omega}, \end{aligned}$$

which is (4.7). Since we also have (with the help of $\|\mathbf{w}^m_\lambda\|_\infty \le C\rho\lambda$)

$$\|\mathbf{w}^m_\lambda\|_{0,p;\,\Omega} \le C\|\mathbf{w}^m\|_{0,p;\,\Omega}$$

and $\mathbf{w}^m \rightharpoonup 0$ weakly in $\mathring{W}^{1,p}(\Omega)$ we use compact embedding and interpolation to conclude $(4.8)_1$. From this and (4.5), equation $(4.8)_2$ follows easily. ∎

Next, we consider $\{(\mathbf{v}^k, P^{I_k})\}^{I=1,2,3,4}_{k\in\mathbb{N}}$ and $(\mathbf{v}, P) \in V_p \times L^{r'}(\Omega)$ satisfying (2.4)–(2.7), (2.10), (2.12)–(2.17), (3.3)–(3.8), and (3.13) and set

$$g^k \equiv C\left(|\mathbf{D}(\mathbf{v}^k)|^p + |\mathbf{D}(\mathbf{v})|^p + |\varphi_2|^{\frac{p}{p-1}} + |P^{1_k}|^{\frac{p}{p-1}}\right), \tag{4.17}$$

where φ_2 comes from (1.6).

Due to a priori estimates we see that g^k satisfy the uniform bound

$$\int_\Omega g^k\, dx \le K. \tag{4.18}$$

PROPOSITION 4.3. *For a given $\varepsilon > 0$ there are a subsequence $\{\mathbf{v}^\ell\}_{\ell\in\mathbb{N}} \subset \{\mathbf{v}^k\}_{k\in\mathbb{N}}$ and $\lambda \ge \frac{1}{\varepsilon}$ independent of ℓ such that*

$$\int_{F^\ell_\lambda} g^\ell\, dx \le \varepsilon, \tag{4.19}$$

where

$$F^\ell_\lambda \equiv \{x \in \Omega; \lambda < M(\nabla(\mathbf{v}^\ell - \mathbf{v}))(x) \le \lambda^2\}. \tag{4.20}$$

Proof of Proposition 4.3. For a given $\varepsilon \in (0,1)$ we find $N \in \mathbb{N}$ such that

$$N\varepsilon > K \qquad (K \text{ from (4.18)}) \tag{4.21}$$

and set

$$\lambda_0 = \frac{1}{\varepsilon}. \tag{4.22}$$

For each $k \in \mathbb{N}$ we introduce sets F_i^k, $i = 0, 1, \dots, N-1$,

$$F_i^k \equiv \{x \in \Omega; \lambda_0^{2^i} < M(\nabla(\mathbf{v}^k - \mathbf{v}))(x) \le \lambda_0^{2^{i+1}}\},$$

which are for fixed k mutually disjoint. Thus, due to (4.18),

$$\sum_{i=0}^{N-1} \int_{F_i^k} g^k \, dx \le K.$$

Due to (4.21), however, for each k there is an index $i(k)$ such that

$$\int_{F_{i(k)}^k} g^k \, dx \le \varepsilon.$$

As $i(k)$'s take values from the finite set $\{0, 1, \dots, N-1\}$ there exists certainly a subsequence $\{\mathbf{v}^\ell\}_{\ell \in \mathbb{N}}$ of $\{\mathbf{v}^k\}_{k \in \mathbb{N}}$ and an index $i_0 \in \{0, 1, \dots, N-1\}$ so that $i(\ell) = i_0$ for all $\ell \in \mathbb{N}$. Then setting $\lambda = \lambda_0^{2^{i_0}}$ and defining F_λ^ℓ as in (4.20) we observe that Proposition 4.3 is proved. □

PROPOSITION 4.4. *Let $\theta \in (0,1)$ be chosen and $\eta > 0$ be arbitrary. Then the sequence $\{\mathbf{v}^\ell\}_{\ell \in \mathbb{N}}$ determined in Proposition* 4.3 *satisfies* (4.1).

Proof of Proposition 4.4. We fix $p \in (\frac{2d}{d+2}, \frac{2d}{d+1}]$ and recall that $\frac{1}{r} = \frac{(d+2)p - 2d}{dp}$. Then we take $\varepsilon > 0$ so small that condition (4.32) specified at the end of the proof is fulfilled. To this ε, find $\{\mathbf{v}^\ell\}_{\ell \in \mathbb{N}} \subset \{\mathbf{v}^k\}_{k \in \mathbb{N}}$ and $\lambda \ge \frac{1}{\varepsilon}$ such that Proposition 4.3 holds. Now, we apply Proposition 4.1 to $(\mathbf{v}^\ell - \mathbf{v})$ and use the Lipschitz truncation $(\mathbf{v}^\ell - \mathbf{v})_\lambda$ as a test function in (3.13). We also subtract from both sides of the obtained equality the term

$$\int_\Omega \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v})) : \mathbf{D}((\mathbf{v}^\ell - \mathbf{v})_\lambda) \, dx.$$

Then we use the facts that $\mathbf{v}^\ell - \mathbf{v} = (\mathbf{v}^\ell - \mathbf{v})_\lambda$ on $\Omega \setminus A_\lambda^\ell$. Thus, $\operatorname{div}(\mathbf{v}^\ell - \mathbf{v})_\lambda = 0$ almost everywhere on $\Omega \setminus A_\lambda^\ell$.

As a result of this consideration we obtain

$$
\begin{aligned}
J^\ell &\equiv \int_{\Omega\setminus A^\ell_\lambda} \Big[\mathbf{T}(\cdot,\mathbf{D}(\mathbf{v}^\ell)) - \mathbf{T}(\cdot,\mathbf{D}(\mathbf{v}))\Big] : \mathbf{D}(\mathbf{v}^\ell - \mathbf{v})\,dx \\
&= \int_{A^\ell_\lambda} \Big[\mathbf{T}(\cdot,\mathbf{D}(\mathbf{v})) - \mathbf{T}(\cdot,\mathbf{D}(\mathbf{v}^\ell))\Big] : \mathbf{D}((\mathbf{v}^\ell - \mathbf{v})_\lambda)\,dx \\
&\quad - \int_{A^\ell_\lambda} P^{1\ell}\mathrm{div}\,((\mathbf{v}^\ell - \mathbf{v})_\lambda)\,dx \\
&\quad - \int_{A^\ell_\lambda} (P^{2\ell} + P^{3\ell})\mathrm{div}\,((\mathbf{v}^\ell - \mathbf{v})_\lambda)\,dx \\
&\quad + \int_\Omega \Big(\mathbf{v}^\ell \otimes (\mathbf{v}^\ell - \mathbf{v}) + (\mathbf{v}^\ell - \mathbf{v}) \otimes \mathbf{v}\Big) : \mathbf{D}((\mathbf{v}^\ell - \mathbf{v})_\lambda)\,dx \\
&\quad + \int_\Omega \Big(\nabla P^{4\ell} - \frac{1}{\ell}|\mathbf{v}^\ell|^{q-2}\mathbf{v}^\ell\Big)\cdot(\mathbf{v}^\ell - \mathbf{v})_\lambda\,dx \\
&\quad + \int_\Omega (\nabla\mathbf{U} - \mathbf{T}(\cdot,\mathbf{D}(\mathbf{v}))) : \nabla(\mathbf{v}^\ell - \mathbf{v})_\lambda\,dx \\
&\equiv I^\ell_1 + I^\ell_2 + I^\ell_3 + I^\ell_4 + I^\ell_5 + I^\ell_6.
\end{aligned}
\tag{4.23}
$$

We evaluate terms on the right-hand side of (4.23) one after another. Note that λ is fixed and

$$
\|\nabla(\mathbf{v}^\ell - \mathbf{v})_\lambda\|_{0,\infty;\,\Omega} \le C\lambda. \tag{4.24}
$$

We are interested in showing that all terms I^ℓ_k, $k = 1,2,3,4,5,6$, are small for $\ell \to \infty$. First, using the compactness (2.14) and (3.7) together with (4.24) we observe that

$$
\lim_{\ell\to\infty} I^\ell_3 + I^\ell_4 = 0. \tag{4.25}
$$

But the same is true for I^ℓ_5 due to (3.8), (2.18), and (4.24). Thus

$$
\lim_{\ell\to\infty} I^\ell_5 = 0. \tag{4.26}
$$

When dealing with I^ℓ_1 and I^ℓ_2 we use Propositions 4.1 and 4.3 and the Hölder inequality

$$
\begin{aligned}
|I^\ell_1 + I^\ell_2| &= \Big|\int_{F^\ell_\lambda\cup G^\ell_\lambda} \Big[\mathbf{T}(\cdot,\mathbf{D}(\mathbf{v})) - \mathbf{T}(\cdot,\mathbf{D}(\mathbf{v}^\ell)) - P^{1\ell}\mathbf{I}\Big] : \nabla(\mathbf{v}^\ell - \mathbf{v})_\lambda\,dx\Big| \\
&\le \int_{F^\ell_\lambda} |\,\ldots\ldots\,|\,dx + \int_{G^\ell_\lambda} |\,\ldots\ldots\,|\,dx \\
&\le \Big(\int_{F^\ell_\lambda} g^\ell\,dx\Big)^{\frac{p-1}{p}} \|\nabla(\mathbf{v}^\ell - \mathbf{v})_\lambda\|_{0,p,F^\ell_\lambda} + C\lambda\Big(\int_{G^\ell_\lambda} g^\ell\,dx\Big)^{\frac{p-1}{p}} |G^\ell_\lambda|^{\frac{1}{p}} \\
&\le K\left(\varepsilon^{1-\frac{1}{p}} + \frac{C}{\lambda}\right) \le K\,C(\varepsilon^{1-\frac{1}{p}} + \varepsilon).
\end{aligned}
\tag{4.27}
$$

Further, from (4.8) applied to $(\mathbf{v}^\ell - \mathbf{v})_\lambda$ we know particularly that

$$
(\mathbf{v}^\ell - \mathbf{v})_\lambda \rightharpoonup 0 \qquad \text{weakly in } \mathring{W}^{1,r}(\Omega). \tag{4.28}
$$

Since $\nabla \mathbf{U} \in L^{r'}(\Omega)$ and $\mathbf{T}(\cdot, \mathbf{D}(\mathbf{v})) \in L^{p'}(\Omega)$ we have

$$\lim_{\ell \to \infty} I_6^\ell = 0. \tag{4.29}$$

To summarize, we have observed that

$$\lim_{\ell \to \infty} J^\ell = KC(\varepsilon^{1-\frac{1}{p}} + \varepsilon). \tag{4.30}$$

Finally, fix $\theta \in (0,1)$ and denote the integral in (4.1) by Y^ℓ. Then we have

$$\begin{aligned} Y^\ell \le & \int_{\Omega \setminus A_\lambda^\ell} \Big[(\mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}^\ell)) - \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}))) : \mathbf{D}(\mathbf{v}^\ell - \mathbf{v}) \Big]^\theta \, dx \\ & + \int_{A_\lambda^\ell} \Big[(\mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}^\ell)) - \mathbf{T}(\cdot, \mathbf{D}(\mathbf{v}))) : \mathbf{D}(\mathbf{v}^\ell - \mathbf{v}) \Big]^\theta \, dx. \end{aligned} \tag{4.31}$$

With help from the Hölder inequality and a priori estimates we obtain

$$Y^\ell \le (J^\ell)^\theta |\Omega \setminus A_\lambda^\ell|^{1-\theta} + K |A_\lambda^\ell|^{1-\theta}.$$

Using (4.31) and (4.9)–(4.11) we finally conclude

$$\begin{aligned} \lim_{\ell \to \infty} Y^\ell & \le |\Omega|^{1-\theta} (K\,C)^\theta (\varepsilon^{1-\frac{1}{p}} + \varepsilon)^\theta + K \left(\frac{C}{\lambda^p} \right)^{1-\theta} \\ & \le |\Omega|^{1-\theta} (K\,C)^\theta (\varepsilon^{1-\frac{1}{p}} + \varepsilon)^\theta + K\,(K\,C)^{1-\theta} \varepsilon^{p(1-\theta)}. \end{aligned}$$

If ε is so small that

$$|\Omega|^{1-\theta} (K\,C)^\theta (\varepsilon^{1-\frac{1}{p}} + \varepsilon)^\theta + K\,(K\,C)^{1-\theta} \varepsilon^{p(1-\theta)} < \eta, \tag{4.32}$$

then Proposition 4.4 and consequently Theorem 1.1 are proved. □

5. Final remarks. (1) The proof of Theorem 1.1 offers also another argument for the existence result in the case $p = \frac{2d}{d+1}$. This limiting case (for this p the convective term is "a priori" only in L^1) was included in our previous existence result in [12] where we used the fact that the convective term $(\mathbf{v} \cdot \nabla)\mathbf{v}$ belongs locally to the Hardy space $\mathcal{H}^1$ (due to $\operatorname{div} \mathbf{v} = 0$) and the duality of $\mathcal{H}^1$ and BMO (= the John–Nirenberg space of functions with bounded mean oscillation). The above given proof of Theorem 1.1 also works (of course) in this case, and therefore we do not need to use the above-mentioned facts (in this case).

(2) On the other hand one can give an alternative proof of Theorem 1.1 by using the following compensated integrability result: For $\frac{1}{r} = \frac{2}{p} - \frac{1}{d} = \frac{2d-p}{dp}$ and $\mathbf{w} \in V_p$ it holds that

$$(\mathbf{w} \cdot \nabla)\mathbf{w} \in h^r(\Omega),$$

where $h^r(\Omega)$ denotes the local Hardy space. Observe that $\frac{d}{d+1} < r < 1$ is equivalent to $\frac{2d}{d+2} < p < \frac{2d}{d+1}$ (see [5], [27]).

Taking [35, section 2.11.3, pp. 180–182; section 2.5.7, pp. 89–91; and section 2.5.12, pp. 109–114] into account we can dispose of

$$h^r(\Omega) \equiv F^0_{r,2}(\Omega) \equiv \text{Triebel–Lizorkin space for } 0 < r \le 1$$

and

$$(h^r(\Omega))' \equiv (F^0_{r,2}(\Omega))' \equiv B^{d(\frac{1}{r}-1)}_{\infty,\infty}(\Omega) \equiv B^{\frac{2d-(d+1)p}{p}}_{\infty,\infty}(\Omega) \equiv C^{0,\alpha}(\overline{\Omega}),$$

where $\alpha = \frac{2d-(d+1)p}{p} \in (0,1)$ if and only if $\frac{2d}{d+2} < p < \frac{2d}{d+1}$. Noticing that our test function $\mathbf{\Phi}^k_\lambda = (\mathbf{v}^k - \mathbf{v})_\lambda$ belongs to $\mathring{W}^{1,\beta}(\Omega)$ for all finite β we certainly have $\mathbf{\Phi}^k_\lambda \in C^{0,\alpha}(\overline{\Omega})$. The only difference to the above given proof appears now in dealing with the convective term: Instead of integrating by parts we keep it in the form

$$\int_\Omega (\mathbf{v}^k \cdot \nabla)\mathbf{v}^k \cdot \mathbf{\Phi}^k_\lambda \, dx \equiv \langle (\mathbf{v}^k \cdot \nabla)\mathbf{v}^k, \mathbf{\Phi}^k_\lambda \rangle,$$

where the brackets now denote the duality between h^r and $C^{0,\alpha}$, use the uniform boundedness of $(\mathbf{v}^k \cdot \nabla)\mathbf{v}^k$ in h^r, and have to ensure that it converges to zero for $k \to \infty$. This can be achieved by observing that

$$\mathbf{\Phi}^k_\lambda \to 0 \quad \text{strongly in } C^{0,\alpha} \qquad \text{for some } \alpha \in (0,1).$$

This follows, however, from $(4.8)_1$ and the interpolation inequalities

$$\begin{aligned}\|\mathbf{\Phi}^k_\lambda\|_{C^{0,\alpha}} &\le C\,\|\mathbf{\Phi}^k_\lambda\|^\theta_\infty\,\|\mathbf{\Phi}^k_\lambda\|^{1-\theta}_{C^{0,\beta}} \\ &\le C\|\mathbf{\Phi}^k_\lambda\|^{\frac{\theta}{2}}_{0,2d}\|\mathbf{\Phi}^k_\lambda\|^{\frac{\theta}{2}}_{1,2d}\,\|\mathbf{\Phi}^k_\lambda\|^{1-\theta}_{C^{0,\beta}} \le C\|\mathbf{\Phi}^k_\lambda\|^{\frac{\theta}{2}}_{0,2d}\|\mathbf{\Phi}^k_\lambda\|^{1-\frac{\theta}{2}}_{1,s},\end{aligned}$$

valid for $0 < \alpha < \beta < 1$, $\theta = 1 - \frac{\alpha}{\beta}$, $1-\theta = \frac{\alpha}{\beta}$, and $s \ge 2d$ so that $W^{1,s} \hookrightarrow C^{0,\beta}$.

The rest of the proof coincides with the original proof.

(3) Using the method of proof of our main theorem we can also generalize the result of Dal Maso and Murat [6] to include "some" nonlinear terms on the right-hand side satisfying "suitable" growth conditions, but we will not follow these possibilities here.

(4) Another possible use of our scheme of proof developed here would be in the theory of electrorheological fluids with shear-dependent viscosities (steady flows), but this will be a future project. The interested reader is referred to [33].

(5) The $C^{1,1}$-regularity of boundary is required in Theorem 1.1 and its proof due to the applications of the L^q-theory for the Stokes system in various places. However, the result of Theorem 1.1 holds for the Lipschitz domains. ($C^{0,1}$-regularity of the boundary seems to be needed to have the global estimates on the pressure.) Considering Lipschitz domains within the proof would certainly reduce the readability of the paper. This is why we finally prefer to deal with domains of the $C^{1,1}$-class.

(6) It follows from the proof that the pressure P corresponding to the weak solution $\mathbf{v} \in V_p$ belongs to $L^{r'}$ where $r' = \frac{dp}{2(d-p)}$.

REFERENCES

[1] E. Acerbi and N. Fusco, *Semicontinuity problems in the calculus of variations*, Arch. Ration. Mech. Anal., 86 (1984), pp. 125–145.

[2] E. Acerbi and N. Fusco, *An approximation lemma for $W^{1,p}$-functions*, in Material Instabilities in Continuum Mechanics and Related Mathematical Problems, J. M. Ball, ed., Oxford University Press, New York, 1998, pp. 1–5.

[3] Ch. Amrouche and V. Girault, *Decomposition of vector spaces and application to the Stokes problem in arbitrary dimension*, Czechoslovak Math. J., 44 (1994), pp. 109–140.

[4] H. Bellout, F. Bloom, and J. Nečas, *Young measure-valued solutions for non-Newtonian incompressible fluids*, Comm. Partial Differential Equations, 19 (1994), pp. 1763–1803.

[5] R. Coifman, P. L. Lions, Y. Meyer, and S. Semmes, *Compensated compactness and Hardy spaces*, J. Math. Pures Appl., 72 (1993), pp. 247–286.
[6] G. Dal Maso and F. Murat, *Almost everywhere convergence of gradients of solutions to nonlinear elliptic systems*, Nonlinear Anal., 31 (1998), pp. 405–412.
[7] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.
[8] G. Dolzmann, N. Hungerbühler, and S. Müller, *Non-linear elliptic systems with measure-valued right hand side*, Math. Z., 226 (1997), pp. 545–574.
[9] G. Dolzmann, N. Hungerbühler, and S. Müller, *Uniqueness and maximal regularity for nonlinear elliptic systems of n-Laplace type with measure valued right hand side*, J. Reine Angew. Math., 520 (2000), pp. 1–35.
[10] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, CRC Press, Boca Raton, FL, 1992.
[11] J. Frehse and J. Málek, *Problems due to the no-slip boundary in incompressible fluid dynamics*, in Geometric Analysis and Nonlinear Partial Differential Equations, S. Hildebrandt and H. Karcher, eds., Springer-Verlag, New York, 2003, pp. 559–571.
[12] J. Frehse, J. Málek, and M. Steinhauer, *An existence result for fluids with shear dependent viscosity—steady flows*, Nonlinear Anal., 30 (1997), pp. 3041–3049.
[13] J. Frehse, J. Málek, and M. Steinhauer, *On existence results for fluids with shear dependent viscosity—unsteady flows*, in Partial Differential Equations, Theory and Numerical Solution, W. Jäger, J. Nečas, O. John, K. Najzar, and J. Stará, eds., Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 121–129.
[14] M. Fuchs and G. A. Seregin, *Some remarks on non-Newtonian fluids including nonconvex perturbations of the Bingham and Powell-Eyring model for visco-plastic fluids*, Math. Models Methods Appl. Sci., 7 (1997), pp. 405–433.
[15] M. Fuchs and G. A. Seregin, *Variational Methods for Problems from Plasticity Theory and for Generalized Newtonian Fluids*, Lecture Notes in Math. 1749, Springer-Verlag, Berlin, Heidelberg, New York, 2000.
[16] L. Greco, T. Iwaniec, and C. Sbordone, *Variational integrals of nearly linear growth*, Differential Integral Equations, 10 (1997), pp. 687–716.
[17] P. Kaplický, J. Málek, and J. Stará, *On global existence of smooth two-dimensional steady flows for a class of non-Newtonian fluids under various boundary conditions*, in Applied Nonlinear Analysis, A. Sequeira, H. Beirao da Veiga, and J. H. Videman, eds., Kluwer Academic/Plenum Publishers, New York, 1999, pp. 213–229.
[18] O. A. Ladyzhenskaya, *On some new equations describing dynamics of incompressible fluids and on global solvability of boundary value problems to these equations*, Trudy Steklov's Math. Institute, 102 (1967), pp. 85–104.
[19] O. A. Ladyzhenskaya, *On some modifications of the Navier-Stokes equations for large gradients of velocity*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI), 7 (1968), pp. 126–154.
[20] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1969.
[21] R. Landes, *Quasimonotone versus pseudomonotone*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 705–717.
[22] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris, 1969.
[23] J. Málek, J. Nečas, and M. Růžička, *On the non–Newtonian incompressible fluids*, Math. Models Methods Appl. Sci., 3 (1993), pp. 35–63.
[24] J. Málek, J. Nečas, and M. Růžička, *On weak solutions to a class of non–Newtonian incompressible fluids in bounded three-dimensional domains. The case $p \geq 2$*, Adv. Differential Equations, 6 (2001), pp. 257–302.
[25] J. Málek, J. Nečas, M. Růžička, and M. Rokyta, *Weak and Measure–valued Solutions to Evolutionary Partial Differential Equations*, Appl. Math. Math. Comput. 13, Chapman & Hall, London, 1996.
[26] J. Málek, K. R. Rajagopal, and M. Růžička, *Existence and regularity of solutions and stability of the rest state for fluids with shear dependent viscosity*, Math. Models Methods Appl. Sci., 6 (1995), pp. 789–812.
[27] S. Müller, *Hardy space methods for nonlinear partial differential equations*, Tatra Mt. Math. Publ., 4 (1994), pp. 159–168.
[28] S. Müller, *A sharp version of Zhang's theorem on truncating sequences of gradients*, Trans. Amer. Math. Soc., 351 (1999), pp. 4585–4597.
[29] J. Nečas, *Sur les normes équivalentes dans $W_p^k(\Omega)$ et sur la coercivité des formes formellement positives*, in Séminaire Equations aux Dérivées Partielles, Montreal, Les presses de

l´Université de Montréal, 1966, pp. 102–128.
[30] M. Pokorný, *Cauchy problem for the non-Newtonian viscous incompressible fluid*, Appl. Math., 41 (1996), pp. 169–201.
[31] K. R. Rajagopal, *Mechanics of non-Newtonian fluids*, in Recent Developments in Theoretical Fluid Mechanics, G. P. Galdi and J. Nečas, eds., Res. Notes in Math. 291, Longman Scientific and Technical, Harlow, UK, 1993, pp. 129–162.
[32] M. Růžička, *A note on steady flow of fluids with shear dependent viscosity*, Nonlinear Anal., 30 (1997), pp. 3029–3039.
[33] M. Růžička, *Electrorheological Fluids: Modeling and Mathematical Theory*, Lecture Notes in Math. 1748, Springer-Verlag, Berlin, 2000.
[34] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[35] H. Triebel, *Theory of Function Spaces*, Birkhäuser-Verlag, Basel, Boston, Stuttgart, 1983.
[36] E. Zeidler, *Nonlinear Functional Analysis and Its Applications. Vol.* 2A: *Linear Monotone Operators*, Springer-Verlag, New York, 1990.
[37] E. Zeidler, *Nonlinear Functional Analysis and Its Applications. Vol.* 2B: *Nonlinear Monotone Operators*, Springer-Verlag, New York, 1990.
[38] W. P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.
[39] K. Zhang, *On the Dirichlet problem for a class of quasilinear elliptic systems of PDEs in divergence form*, in Partial Differential Equations, Proc. Tranjin, S. S. Chern, ed., Lecture Notes in Math. 1036, Springer-Verlag, Berlin, 1988, pp. 262–277.
[40] K. Zhang, *Biting theorems for Jacobians and their applications*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 345–365.
[41] K. Zhang, *Remarks on perturbed systems with critical growth*, Nonlinear Anal., 18 (1992), pp. 1167–1179.
[42] K. Zhang, *A construction of quasiconvex functions with linear growth at infinity*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 19 (1992), pp. 313–326.

SIAM J. MATH. ANAL.
Vol. 34, No. 5, pp. 1084–1104

MINIMIZATION PROBLEMS AND ASSOCIATED FLOWS RELATED TO WEIGHTED p ENERGY AND TOTAL VARIATION*

YUNMEI CHEN† AND MURALI RAO†

Abstract. Motivated by the problem of edge preserving regularization in image restoration, in this paper we investigate the relations between weighted p energy based and total variation based minimization problems and their associated flows. We prove that the weighted total variation based minimization and its associated flow in a weakened formulation can be approximated by the weighted p energy based minimization and its associated flows, respectively.

Moreover, we show that the flow of the weighted total variation based minimization converges weakly in BV and strongly in L^2 to the minimizer as $t \to \infty$.

Key words. BV-space, pseudosolutions, trace, minimization, degenerate parabolic equations

AMS subject classifications. 49J40, 35K65

PII. S0036141002404577

1. Introduction.

In this paper, we investigate the relation of two variational models and their associated flows applied in image recovery. One of them is weighted total variation based minimization, and the other one is based on minimizing the weighted L^p-norm of image gradients.

The image recovery problem we consider here is a problem of recovering an unknown data u from an observed noisy data I, which is related to u by $I = u + noise$. The challenging aspect of solving this problem is to selectively filter the noise without losing significant features. Standard regularization methods, for instance, Gaussian filters corresponding to minimizing L^2-norm of image gradients, cannot solve this problem since they are isotropical smoothing and do not allow discontinuous solutions. One approach for solving this problem is to use total variation based regularization—that is, to minimize the total variation norm of u subject to the constraint $\|u - I\|_{L^2}^2 = \sigma^2$ (see [15], [16]), or to solve the related unconstrained minimization problem

$$\min_u \int_\Omega |\nabla u| + \frac{\lambda}{2}|u - I|^2$$

(see [1], [4], [24], and the references therein). To smooth images more selectively, spatially adaptive total variation models were proposed (see, e.g., [17], [18], [19], [3]), where the adaptivity is realized by using a weighted total variation norm. These models can be represented in the following framework:

$$\min_u E(u) =: \min_u \int_\Omega \alpha(x)|\nabla u| + \int_\Omega \frac{1}{2}|u - I|^2 dx,$$

where $\alpha(x)$ is chosen to be inversely proportional to the likelihood of the presence of an edge. Certain choices of $\alpha(x)$ based on image features or noise levels or both were given in [17], [18], and [19]. Their numerical results showed the effectiveness of the model in removing noise and sharpening image features.

*Received by the editors March 27, 2002; accepted for publication (in revised form) September 6, 2002; published electronically April 15, 2003.

http://www.siam.org/journals/sima/34-5/40457.html

†Department of Mathematics, University of Florida, 358 Little Hall, Gainesville, FL (yun@math.ufl.edu, murali@ufl.edu). The first author's research was partially supported by DMS-9972662 and NIH P50-DC03888.

On the other hand, minimizing L^p-norms $(1 < p < 2)$ of the gradient has been employed as one of the approximation methods for image denoising and has been considered as an alternate way between a total variation based approach and one which minimizes the L^2-norm of the gradient depending on the needs of a particular image [2]. More applications of minimizing L^p-norms $(1 < p < 2)$ of the gradient in image processing can be found in [23].

Motivated by these works, we are interested in the mathematical theories on the relations of the solutions to the following two minimization problems and their associated flows:

$$\min_v E(v) =: \min_{v \in BV_g \cap L^2} \int_\Omega \alpha(x)|\nabla v| + \frac{1}{2}\int_\Omega |v - I|^2 dx \tag{1.1}$$

and

$$\min_v E_p(v) =: \min_{v \in W_g^{1,p} \cap L^2} \frac{1}{p}\int_\Omega \alpha(x)|\nabla v|^p + \frac{1}{2}|v - I|^2 dx, \tag{1.2}$$

where

$$BV_g = \{v \in BV(\Omega) | v = g \text{ on } \partial\Omega\}, \qquad W_g^{1,p} = \{v \in W^{1,p}(\Omega) | v = g \text{ on } \partial\Omega\}.$$

Here the boundary conditions are in the sense of trace. The purpose of this paper is to investigate if the solution of (1.1) and its associated flow can be approximated by a sequence of solutions to (1.2) and their associated flows. This study is of more than theoretical interest because it will provide an alternate approximation scheme for solving total variation based minimization problems and their associated flows that have many applications.

The existence of a Lipschitz continuous solution to the problem $\min_{u \in BV_g} \int_\Omega |\nabla u|$ has been obtained under certain conditions on Ω and g (see [12], [13], [14], [20], [21], and [26]). However, it seems difficult to apply their arguments to the problem (1.1) due to the presence of $\alpha(x)$ and the term $|v - I|^2$ in $E(v)$.

Although much attention to the existence and regularity for the solutions to the problem $\min \int_\Omega |\nabla u|^p dx$ with fixed boundary data has been given (see [7], [8], [9], and references therein), we found, after careful checking, that the Lipschitz continuous solutions obtained in pioneering works have their Lipschitz constants depending on p and tending to infinity as $p \to 1$.

The existence and uniqueness for the solution of (1.1) in bounded variation (BV-) space for constant α has been proved in [1] and [4]. Results for more general cases were obtained by Vese in [25] for the functional

$$F(u) = \int_\Omega (Ku - u_0)^2 dx + \alpha \int_\Omega \phi(|\nabla u|)$$

and its associated flow. Here $\alpha \geq 0$ is constant, $\phi : \mathbf{R} \to \mathbf{R}^+$ is a convex, even function nondecreasing in $\mathbf{R}^+$ with linear growth, and $K : L^p(\Omega) \to L^2(\Omega)$ is a linear, continuous, and injective operator. The existence result for the flow associated with this minimization problem is only in dimensions one and two because the methods employed there use general results on maximal monotone operators and evolution operators in Hilbert spaces.

The existence, uniqueness, and large time asymptotic behavior for the flow associated with the minimization problem

$$\min_{u \in BV_g} \int_\Omega \phi(Du),$$

i.e.,

$$\frac{\partial u}{\partial t} = \mathrm{div}_x(\phi_p(\nabla u)) \quad \text{in } \Omega \times R_+,$$

$$u = g \ \text{ on } \ \partial\Omega \times R_+,$$

$$u = u_0 \ \text{ on } \Omega \times \{0\},$$

has been proved in [10], provided that ϕ is a convex linear-growth function. A typical example of ϕ is

$$\phi(q) =: 1/2|q|^2 \text{ if } |q| \le 1, \text{ and } \phi(q) =: |q| - 1/2 \text{ if } |q| \ge 1;$$

the same results have been shown in [26]. The key idea of their work is to approximate the solution to the equation $\frac{\partial u}{\partial t} = \mathrm{div}_x(\phi_p(\nabla u))$ by the solutions to the equations $\frac{\partial u}{\partial t} = \mathrm{div}_x(\phi_p^\epsilon(\nabla u))$ with the same initial and boundary conditions, where $\phi_\epsilon(q) = \eta_\epsilon * \phi(q)$ and η_ϵ is the standard molifier.

In this paper, we will use the so-called *relaxed energy* (see [10], [22], and [26]) to define a pseudosolution u of (1.1) as a solution to the problem

$$\min_{v \in BV(\Omega) \cap L^2} E^g(v),$$

where

$$E^g(v) =: \int_\Omega \alpha(x)|\nabla v| + \frac{1}{2}|v - I|^2 dx + \int_{\partial\Omega} \alpha(x)|v - g| dH^{n-1}.$$

We will show that the pseudosolution of (1.1) can be approximated weakly as $p \downarrow 1$ either by the solutions to (1.2) or by a sequence of functions that minimize the functional

$$E_p^g(v) =: \frac{1}{p}\int_\Omega \alpha(x)|\nabla v|^p + \frac{1}{2}|v - I|^2 dx + \int_{\partial\Omega} \alpha(x)|v - g| dH^{n-1}$$

over $v \in W^{1,p}(\Omega) \cap L^2$.

Moreover, we will show the existence and uniqueness of a pseudosolution $u(x,t)$ to the flow associated with (1.1):

$$\partial_t u - \mathrm{div}(\alpha(x)\nabla u/|\nabla u|) + (u - I) = 0, \quad x \in \Omega, \ t > 0, \tag{1.3}$$

$$u(x,0) = I(x), \quad x \in \Omega, \tag{1.4}$$

$$u(x,t) = g(x), \quad x \in \partial\Omega, \ t \ge 0. \tag{1.5}$$

This will be done by proving that the weak solutions to the flow associated with (1.2) converge weakly in $BV \cap L^2$ as $p \to 1$ to a pseudosolution of (1.3)–(1.5).

Finally, we prove that as $t \to \infty$ the solution $u(\cdot, t)$ of (1.3)–(1.5) converges strongly in L^2 and weakly in BV to the solution of the relaxed problem associated with (1.1).

We would like to point out that all the proofs presented in this paper can easily be carried over to the problem of minimizing the energy function in (1.1) over $BV(\Omega) \cap L^2$ and the flow (1.3)–(1.4) with a free Neumann boundary condition by just dropping the boundary term.

2. Preliminaries. From now on we always assume the following:
(H.1) Ω is a bounded open subset of $\mathbf{R}^n$ with Lipschitz boundary.
(H.2) $\alpha(x)$ is a positive valued continuous function on R^n.
In practical image restoration problems, $\alpha(x)$ may be chosen as

$$\alpha(x) = \frac{\alpha_1}{1 + k|\nabla G_\sigma * I|^2}, \tag{2.1}$$

where $G_\sigma(x) = \frac{1}{\sigma}\exp(-|x|^2/4\sigma^2)$ and $\alpha_1 > 0$, $k > 0$, and $\sigma > 0$ are parameters. With this choice, $\alpha(x)$ satisfies assumption (H.2) and takes much smaller values near likely edges than on homogeneous regions so that edges are much less smoothed and, hence, preserved. Moreover, if $I \in L^\infty(\Omega)$, then

$$0 < \alpha_0 \le \alpha(x) \le \alpha_1, \quad x \in \Omega, \tag{2.2}$$

with $\alpha_0 = \alpha_1/(1 + C(k,\sigma)|I|_{L^\infty(\Omega)})$.

Now we define the weighted total variation norm with weight function α satisfying (H.2).

DEFINITION 2.1. *A function $f \in L^1(\Omega)$ has bounded α-total variation in Ω if*

$$\sup_{\phi \in \Phi(\alpha,\Omega)} \int_\Omega f \mathrm{div}\phi dx < \infty,$$

where

$$\Phi(\alpha,\Omega) =: \{\phi \in C_0^1(\Omega, R^n)|\ |\phi| \le \alpha\}. \tag{2.3}$$

We can see that if $f \in L^1(\Omega)$ has bounded α-total variation in Ω, there is a Radon vector measure ∇f on Ω such that

$$\int_\Omega \alpha|\nabla f| =: \sup_{\phi \in \Phi_\alpha} \int_\Omega f \mathrm{div}\phi dx. \tag{2.4}$$

Under assumption (H.2), $f \in L^1(\Omega)$ having bounded α-total variation in Ω implies that $f \in BV(\Omega)$.

We further assume the following:
(H.3) $I \in L^2(\Omega)$, and g is the trace of a function $G \in H^1(\Omega)$.
Under assumptions (H.1)–(H.3), for $v \in BV(\Omega) \cap L^2$, we define

$$E^g(v) =: \int_\Omega \alpha(x)|\nabla v| + \frac{1}{2}|v - I|^2 dx + \int_{\partial\Omega} \alpha(x)|v - g| dH^{n-1}, \tag{2.5}$$

and, for $v \in W^{1,p}(\Omega) \cap L^2$, we define

$$E_p^g(v) =: \frac{1}{p}\int_\Omega \alpha(x)|\nabla v|^p + \frac{1}{2}|v - I|^2 dx + \int_{\partial\Omega} \alpha(x)|v - g| dH^{n-1}. \tag{2.6}$$

PROPOSITION 2.2. *Let* (H.1)–(H.2) *hold. Assume that $f_1 \in BV(\Omega)$ and $f_2 \in BV(R^n - \bar{\Omega})$. Define*

$$\bar{f} = \begin{cases} f_1 & \text{if } x \in \Omega, \\ f_2 & \text{if } x \in R^n - \bar{\Omega}. \end{cases}$$

Then $\bar{f} \in BV(R^n)$, and

$$\int_{R^n} \alpha(x)|\nabla \bar{f}| = \int_\Omega \alpha(x)|\nabla f_1| + \int_{R^n-\bar{\Omega}} \alpha(x)|\nabla f_2| + \int_{\partial\Omega} \alpha(x)|Tf_1 - Tf_2| dH^{n-1}, \tag{2.7}$$

where Tf denotes the trace of $f \in BV(\Omega)$ on $\partial\Omega$.

Proof. For each $\phi \in C_0^1(R^n; R^n)$, $|\phi(x)| \le \alpha(x)$ on Ω, denoting μ to be the unit outward normal to Ω, we have

$$\int_{R^n} \bar{f} \mathrm{div}\phi dx = \int_\Omega f_1 \mathrm{div}\phi dx + \int_{R^n-\bar{\Omega}} f_2 \mathrm{div}\phi dx$$

$$= -\int_\Omega \phi \cdot \nabla f_1 - \int_{R^n-\bar{\Omega}} \phi \cdot \nabla f_2 + \int_{\partial\Omega} (Tf_1 - Tf_2)(\phi \cdot \mu) dH^{n-1}$$

$$\le \int_\Omega \alpha|\nabla f_1| + \int_{R^n-\bar{\Omega}} \alpha|\nabla f_2| + \int_{\partial\Omega} \alpha|Tf_1 - Tf_2| dH^{n-1}.$$

Therefore, $\int_{R^n} \alpha|\nabla \bar{f}| < \infty$, and then $\bar{f} \in BV(R^n)$. Moreover, for each $\phi \in C_0^1(R^n; R^n)$,

$$\int_{R^n} \phi \cdot \nabla \bar{f} = \int_\Omega \phi \cdot \nabla f_1 + \int_{R^n-\bar{\Omega}} \phi \cdot \nabla f_2 - \int_{\partial\Omega} (Tf_1 - Tf_2)(\phi \cdot \mu) dH^{n-1}.$$

Therefore,

$$\nabla \bar{f} = \begin{cases} \nabla f_1 & \text{on } \Omega, \\ \nabla f_2 & \text{on } R^n - \bar{\Omega}. \end{cases}$$

This implies

$$-\int_{\partial\Omega} \phi \cdot \nabla \bar{f} = \int_{\partial\Omega} (Tf_1 - Tf_2)(\phi \cdot \mu) dH^{n-1}.$$

Then

$$\int_{\partial\Omega} \alpha|\nabla \bar{f}| = \int_{\partial\Omega} \alpha|Tf_1 - Tf_2| dH^{n-1}. \tag{2.8}$$

Now (2.7) follows from (2.8). □

THEOREM 2.3. *Let* (H.1)–(H.3) *hold. Assume that $\{f_k\}_{k=1}^\infty \subset BV(\Omega)$ and $f_k \to f$ in $L^1(\Omega)$. Then $f \in BV(\Omega)$, and*

$$\int_\Omega \alpha|\nabla f| + \int_{\partial\Omega} \alpha|f - g| dH^{n-1} \le \liminf_{k\to\infty} \left\{ \int_\Omega \alpha|\nabla f_k| + \int_{\partial\Omega} \alpha|f_k - g| dH^{n-1} \right\}. \tag{2.9}$$

Proof. Let EG be the extension of G such that $EG \in H^1(R^n)$,

$$EG = G \text{ on } \Omega, \quad T(EG) = TG = g \text{ on } \partial\Omega, \quad \text{and} \quad \|EG\|_{H^1(R^n)} \le C\|G\|_{H^1(\Omega)}.$$

Define

$$\bar{f}_k = \begin{cases} f_k & \text{on } \Omega, \\ EG & \text{on } R^n - \bar{\Omega} \end{cases}$$

and

$$\bar{f} = \begin{cases} f & \text{on } \Omega, \\ EG & \text{on } R^n - \bar{\Omega}. \end{cases}$$

Then $\bar{f}_k \to \bar{f}$ in $L^1(R^n)$, and, for each $\phi \in \Phi(\alpha, R^n)$ (see (2.3)),

$$\int_{R^n} \bar{f} \mathrm{div}\phi = \lim_{k\to\infty} \int_{R^n} \bar{f}_k \mathrm{div}\phi \le \liminf_{k\to\infty} \int_{R^n} \alpha |\nabla \bar{f}_k|.$$

Taking the supremum with respect to ϕ over $\Phi(\alpha, R^n)$ yields that

$$\int_{R^n} \alpha |\nabla \bar{f}| \le \liminf_{k\to\infty} \int_{R^n} \alpha |\nabla \bar{f}_k|. \tag{2.10}$$

Moreover, by Proposition 2.2, $\bar{f}_k \in BV(R^n)$ $(k = 1, 2, \ldots)$, and $\bar{f} \in BV(R^n)$, we get

$$\int_{R^n} \alpha |\nabla \bar{f}_k| = \int_{\Omega} \alpha |\nabla f_k| + \int_{R^n - \bar{\Omega}} \alpha |\nabla G| + \int_{\partial\Omega} \alpha |f_k - g| dH^{n-1}, \tag{2.11}$$

$$\int_{R^n} \alpha |\nabla \bar{f}| = \int_{\Omega} \alpha |\nabla f| + \int_{R^n - \bar{\Omega}} \alpha |\nabla G| + \int_{\partial\Omega} \alpha |f - g| dH^{n-1}. \tag{2.12}$$

Passing to the limit $k \to \infty$ in (2.11) and using (2.10) and (2.12), we get (2.9). □

THEOREM 2.4. *Let* (H.1)–(H.3) *hold and* $u \in BV(\Omega) \cap L^2$. *Then, for each* $\epsilon > 0$, *there exists a function* $u_\epsilon \in C^\infty(\bar{\Omega})$ *such that*

$$E^g(u_\epsilon) \le E^g(u) + \epsilon. \tag{2.13}$$

Proof. (1) By a minor modification of the proof of Theorem 1.17 and Remark 1.18 in [11], we can find a sequence $u_j \in C^\infty(\Omega) \cap W^{1,1} \cap L^2$ such that

$$T_r u_j = T_r u \quad \text{in } L^1(\partial\Omega), \tag{2.14}$$

and, as $j \to \infty$,

$$\|u_j - u\|_{L^2(\Omega)} \to 0, \tag{2.15}$$

and

$$\int_{\Omega} |\nabla u_j| \to \int_{\Omega} |\nabla u|. \tag{2.16}$$

Moreover, from (2.16) we have

$$\int_{\Omega} \alpha |\nabla u_j| \to \int_{\Omega} \alpha |\nabla u|. \tag{2.17}$$

The combination of (2.14), (2.15), and (2.17) leads to

$$\lim_{j\to\infty} E^g(u_j) = E^g(u). \tag{2.18}$$

Therefore, for each given $\epsilon > 0$, there exists a function $v_\epsilon \in C^\infty(\Omega) \cap W^{1,1} \cap L^2$ such that

$$E^g(v_\epsilon) \le E^g(u) + \epsilon/2. \tag{2.19}$$

(2) Since $C^\infty(\bar\Omega)$ is dense in $W^{1,1}(\Omega)\cap L^2$, for v_ϵ obtained above, there is a function $u_\epsilon\in C^\infty(\bar\Omega)$ such that

$$\|v_\epsilon-u_\epsilon\|_{W^{1,1}(\Omega)}+\|v_\epsilon-u_\epsilon\|^2_{L^2(\Omega)}$$

$$+\|T_rv_\epsilon-T_ru_\epsilon\|_{L^1(\partial\Omega)}\le\frac{\epsilon}{4\max_{\bar\Omega}\alpha(x)};\tag{2.20}$$

here we have used the trace theorem for the third term in (2.20). By using (H.2), we get from (2.20) that

$$E^g(u_\epsilon)\le E^g(v_\epsilon)+\epsilon/2.\tag{2.21}$$

Now (2.13) follows from (2.19) and (2.21). □

THEOREM 2.5. *Let* (H.1)–(H.3) *hold. Assume that* $u\in BV(\Omega)\cap L^2$, *with* $T_ru=g$ *on* $\partial\Omega$, *where* g *is the trace of a function* $G\in H^1(\Omega)$ *(see* (A.3)*). Then, for each* $\epsilon>0$, *there exists a function* $u_\epsilon\in H^1(\Omega)$ *such that*

$$u_\epsilon=g \text{ on } \partial\Omega \text{ in the sense of trace},\tag{2.22}$$

$$\|u_\epsilon-u\|_{L^2(\Omega)}\le\epsilon,\tag{2.23}$$

$$\int_\Omega\alpha|\nabla u_\epsilon|\le\int_\Omega\alpha|\nabla u|+\epsilon.\tag{2.24}$$

Proof. To see this, we need only to modify the second part of the proof of the above theorem. Let v_ϵ be the function obtained in the proof of the first step. Then v_ϵ verifies (2.14)–(2.16), and $v_\epsilon-G\in W_0^{1,1}(\Omega)\cap L^2$. Therefore, there exists a function $h_\epsilon\in C_0^\infty(\Omega)$ such that

$$\int_\Omega\alpha|\nabla(v_\epsilon-G-h_\epsilon)|+\|v_\epsilon-G-h_\epsilon\|_{L^2(\Omega)}\le\epsilon.$$

Let $u_\epsilon=G+h_\epsilon$. Then $u_\epsilon\in H^1(\Omega)$ satisfies (2.22)–(2.24). □

3. Minimization problems (1.1) and (1.2). In this section, we shall show the existence of the solution to the relaxed problem associated with (1.1) and the convergence of the solutions to the problem (1.2) or to the problem (3.2) below to the pseudosolution of (1.1), as $p\to1$.

Note that the weak limit in $BV(\Omega)$ of a minimizing sequence of (1.1) may fail to have the same boundary data g; hence it may not be in BV_g. As in [10], [22], and [26], we will incorporate the boundary condition into the functional by using the so-called *relaxed energy* and define a pseudosolution of (1.1) as follows.

DEFINITION 3.1. *A function* $u\in BV(\Omega)\cap L^2$ *is a pseudosolution of* (1.1) *if it is a solution to the minimization problem*

$$\min_{v\in BV(\Omega)\cap L^2}E^g(v),\tag{3.1}$$

where $E^g(v)$ *is as defined in* (2.5).

In the appendix, we shall show (see Theorem A.3)

$$\inf_{v\in BV(\Omega)\cap L^2} E^g(v) = \inf_{v\in BV_g\cap L^2} E(v).$$

This is why the solution of (3.1) is called a pseudosolution of (1.1).

Now we prove the existence and uniqueness results for (3.1).

THEOREM 3.2 (existence for (3.1)). *Let* (H.1)–(H.3) *hold. Then there is a unique solution to the problem* (3.1).

Proof. Let $u_n \subset BV(\Omega)\cap L^2$ be a minimizing sequence for E^g; i.e.,

$$E^g(u_n) \to \inf_{v\in BV(\Omega)\cap L^2} E^g(v).$$

By (H.2), α has a lower bound $\alpha_0 > 0$; hence the $u_n's$ are bounded in $BV(\Omega)$ and $L^2(\Omega)$. By the compactness result (see [11] or [5]), there exist a subsequence u_{n_j} of u_n and a function $u \in BV(\Omega)\cap L^2$ such that, as $j\to\infty$,

$$u_{n_j} \to u \text{ strongly in } L^1(\Omega), \quad \text{weakly in } L^2(\Omega).$$

By Theorem 2.3 and the weak lower semicontinuity of the L^2-norm, we get

$$E^g(u) \le \liminf_{j\to\infty} E^g(u_{n_j}) = \inf_{v\in BV(\Omega)\cap L^2} E^g(v).$$

The uniqueness of the minimizer follows from the strict convexity of E^g. ∎

Similarly, we can prove the following theorem.

THEOREM 3.3. *Let* (H.1)–(H.3) *hold. Then there is a unique solution to the problem*

$$\min_{v\in W^{1,p}\cap L^2} E_p^g(v), \tag{3.2}$$

where $E_p^g(v)$ *is as defined in* (2.6).

We also have the following theorem.

THEOREM 3.4 (existence for (1.2)). *Let* (H.1)–(H.3) *hold. Then there is a unique solution to the problem* (1.2).

Proof. Let $u_n \subset W_g^{1,p}\cap L^2$ be a minimizing sequence for E_p. Then the $u_n^{'}s$ are bounded in $W^{1,p}(\Omega)$ and $L^2(\Omega)$ since α is bounded below. Therefore, there exist a subsequence u_{n_j} of u_n and a function $u \in W^{1,p}(\Omega)\cap L^2$ such that, as $j\to\infty$,

$$u_{n_j} \to u \text{ weakly in } W^{1,p}(\Omega) \text{ and } L^2(\Omega), \tag{3.3}$$

and

$$Tu_{n_j} \to Tu \text{ weakly in } L^p(\partial\Omega).$$

Since $Tu_{n_j} = g$ for all j, $Tu = g$ on $\partial\Omega$. Hence $u \in W_g^{1,p}\cap L^2$. From (3.3) and the weak lower semicontinuity of the norms, we get

$$E_p(u) \le \liminf_{j\to\infty} E_p(u_{n_j}) = \inf_{v\in W_g^{1,p}\cap L^2} E_p(v).$$

This shows that u is a solution of (1.2). The uniqueness of the minimizer follows from the strict convexity of E_p. ∎

Next we shall show that the solution of (3.1) can be approximated by both the solutions of (3.2) and (1.2), respectively.

THEOREM 3.5. *Let* (H.1)–(H.3) *hold. Assume* u *is a solution of* (3.1) *and* u_p *are the solutions of* (3.2) *for* $1 < p < 2$. *Then there exists a sequence* u_{p_j} *from* u_p *such that, as* $p_j \to 1$,

$$u_{p_j} \to u \quad \textit{weakly in } BV(\Omega) \textit{ and } L^2(\Omega),$$

and

$$E^g_{p_j}(u_{p_j}) \to E^g(u). \tag{3.4}$$

Proof. Since u_p are solutions of (3.2) for $1 < p < 2$ and α is bounded below by a positive constant, the u_p are bounded in $W^{1,1}(\Omega) \cap L^2$. Therefore, there exist a sequence u_{p_j} from u_p and a function $u_1 \in BV(\Omega) \cap L^2$ such that, as $p_j \to 1$,

$$u_{p_j} \to u_1 \ \text{ strongly in } L^1(\Omega) \text{ and weakly in } L^2(\Omega), \tag{3.5}$$

and ∇u_{p_j} converges to ∇u_1 in the sense of measure. By Theorem 2.4, for u_1, any $\epsilon > 0$, and any fixed $p > 1$, there is a function $u_\epsilon \in W^{1,p}(\Omega) \cap L^2$ such that

$$E^g(u_\epsilon) \le E^g(u_1) + \epsilon. \tag{3.6}$$

Note that

$$E^g_{p_j}(u_{p_j}) = \min_{v \in W^{1,p_j} \cap L^2} E^g_{p_j}(v).$$

From (3.6), we have that

$$\liminf_{p_j \to 1} E^g_{p_j}(u_{p_j}) \le \lim_{p_j \to 1} E^g_{p_j}(u_\epsilon) = E^g(u_\epsilon) \le E^g(u_1) + \epsilon.$$

Letting $\epsilon \to 0$ in the above inequality yields that

$$\lim_{p_j \to 1} E^g_{p_j}(u_{p_j}) \le E^g(u_1). \tag{3.7}$$

On the other hand, by Theorem 2.3 and (3.5),

$$E^g(u_1) \le \liminf_{p_j \to 1} E^g(u_{p_j}) \le \lim_{p_j \to 1} \left\{ \left(\int_\Omega \alpha |\nabla u_{p_j}|^{p_j} dx \right)^{\frac{1}{p_j}} \left(\int_\Omega \alpha dx \right)^{1 - \frac{1}{p_j}} \right.$$

$$\left. + 1/2 \int_\Omega |u_{p_j} - I|^2 dx + \int_{\partial\Omega} \alpha |u_{p_j} - g| dH^{n-1} \right\} = \liminf_{p_j \to 1} E^g_{p_j}(u_{p_j}). \tag{3.8}$$

The combination of (3.7) and (3.8) leads to

$$E^g(u_1) = \lim_{p_j \to 1} E^g_{p_j}(u_{p_j}). \tag{3.9}$$

Next we shall show that u_1 solves (3.1), and hence $u_1 = u$. To see this, let $v \in BV(\Omega) \cap L^2$, and apply Theorem 2.4 to v. Then, for each $\epsilon > 0$, we have a function $v_\epsilon \in W^{1,p}(\Omega) \cap L^2$ such that

$$E^g(v_\epsilon) \le E^g(v) + \epsilon. \tag{3.10}$$

On the other hand, by (3.9) and (3.6),

$$E^g(u_1) = \lim_{p_j \to 1} E^g_{p_j}(u_{p_j}) \le \lim_{p_j \to 1} E^g_{p_j}(v_\epsilon) = E^g(v_\epsilon).$$

Then, from (3.10), we get

$$E^g(u_1) \le E^g(v) + \epsilon. \tag{3.11}$$

Letting $\epsilon \to 0$ in (3.11), we get

$$E^g(u_1) \le E^g(v)$$

for any $v \in BV(\Omega) \cap L^2$. Thus u_1 solves (3.1). By the uniqueness of the solution to (3.1), $u = u_1$, and then, from (3.9),

$$E^g(u) = \lim_{p_j \to 1} E^g_{p_j}(u_{p_j}). \qquad \square$$

THEOREM 3.6. *Let* (H.1)–(H.3) *hold. Assume* u *is a solution of* (3.1) *and* u_p *are the solutions of* (1.2) *for* $1 < p < 2$. *Then there exists a sequence* u_{p_j} *from* u_p *such that, as* $p_j \to 1$,

$$u_{p_j} \to u \quad \textit{weakly in } BV(\Omega) \textit{ and } L^2(\Omega),$$

and

$$E_{p_j}(u_{p_j}) \to E^g(u). \tag{3.12}$$

Proof. Since u_p are solutions of (1.2) for $1 < p < 2$ and α is bounded below by a positive constant, the u_p are bounded in $W^{1,1}(\Omega) \cap L^2$. Therefore, there exist a sequence u_{p_j} from u_p and a function $u_1 \in BV(\Omega) \cap L^2$ such that, as $p_j \to 1$,

$$u_{p_j} \to u_1 \ \text{ strongly in } L^1(\Omega) \text{ and weakly in } L^2(\Omega).$$

Using Theorem 2.3 and noticing the fact that $u_{p_j} = g$ a.e. on $\partial\Omega$, we have

$$E^g(u_1) \le \liminf_{p_j \to 1} E^g_{p_j}(u_{p_j}) = \liminf_{p_j \to 1} E_{p_j}(u_{p_j}). \tag{3.13}$$

Since u_{p_j} is the solution of (1.2) with $p = p_j$, $E^g_{p_j}(u_{p_j}) = \min_{v \in W_g^{1,p_j} \cap L^2} E^g_{p_j}(v)$. By Proposition A.2 and Theorem A.3, we get

$$\liminf_{p_j \to 1} E_{p_j}(u_{p_j}) = \inf_{v \in BV_g \cap L^2} E(v)$$

$$= \inf_{v \in BV(\Omega) \cap L^2} E^g(v) = E^g(u). \tag{3.14}$$

The combination of (3.13)–(3.14) shows that

$$E^g(u_1) \le E^g(u).$$

Therefore, $u_1 = u$, and (3.12) follows from (3.14). $\square$

4. Evolution problems. In this section, we shall show that the flow associated with the minimization problem (1.1), i.e., (1.3)–(1.5), in a weakened formulation can be approximated by a sequence of the solutions to the flows associated with the minimization problem (1.2). We shall also discuss the large time behavior of the solution to (1.3)–(1.5) by an approach different from the subdifferential theory used in both [10] and [25].

Denote $\Omega^T = \Omega \times [0,T)$ and $\partial\Omega^T = \partial\Omega \times [0,T)$, $0 < T \le \infty$.

4.1. Definition of a pseudosolution to (1.3)–(1.5). As mentioned in the previous section, the weak limit (as $p \to 1$) in $BV(\Omega)$ of a sequence in W_g^{1,p_j} may fail to have the same boundary data g. Hence we consider a weakened formulation for the solution to (1.3)–(1.5).

Assume that the solution $u(x,t)$ of (1.3)–(1.5) is sufficiently smooth to justify the following calculations. For arbitrary $v \in L^2(0,T;H^1(\Omega))$, we multiply (1.3) by $v-u$. After integrating by parts and using the convexity of the functions $p \to |p|$ and $u \to |u-I|^2$ and the inequality $\frac{\nabla u}{|\nabla u|} \cdot \mu \le 1$, where μ denotes the exterior unit normal vector to Ω, we get that, for any $t \in [0,T]$,

$$\int_\Omega (\partial_t u)(v-u)dx + E^g(v) \ge E^g(u). \tag{4.1}$$

Then we integrate with respect to t to get that, for any $s \in [0,T]$,

$$\int_0^s \int_\Omega (\partial_t u)(v-u)dxdt + \int_0^s E^g(v)dt \ge \int_0^s E^g(u). \tag{4.2}$$

Conversely, choosing $v = v_\epsilon = u + \epsilon w$, where $w \in C_0^\infty(\Omega)$, in (4.2), and noticing that the left-hand side (LHS) of (4.2) has a minimum at $\epsilon = 0$, we can show that u is a solution of (1.3) in the sense of distribution by computing $\frac{d}{d\epsilon}\int_\Omega (\partial_t(u\epsilon w))dx + E^g(u+\epsilon w)$ at $\epsilon = 0$.

Prompted by these facts, we give the following definition for a pseudosolution of (1.3)–(1.5).

DEFINITION 4.1. *A function $u \in L^2(0,T;BV(\Omega)\cap L^2)$ $(0 < T \le \infty)$ is called a pseudosolution of* (1.3)–(1.5) *if $\partial_t u \in L^2(\Omega^T)$, $u(x,0) = u_0(x)$ on Ω, and u satisfies* (4.2) *for every $v \in L^2([0,T];BV(\Omega)\cap L^2)$, and $s \in [0,T]$.*

Let (H.1)–(H.3) hold. Assume $g \in L^\infty(\partial\Omega)$ and $I \in BV(\Omega)\cap L^\infty$ with $I|_{\partial\Omega} = g$.

We first consider the *approximating* problem

$$\frac{\partial u_{p,\delta}}{\partial t} - \operatorname{div}(\alpha|\nabla u_{p,\delta}|^{p-2}\nabla u_{p,\delta} + (u_{p,\delta} - I_\delta)) = 0 \ \text{ in } \Omega \times \mathbf{R}_+, \tag{4.3}$$

$$u_{p,\delta} = g \ \text{ on } \ \partial\Omega \times \mathbf{R}_+, \tag{4.4}$$

$$u_{p,\delta}(x,0) = I_\delta \ \text{ on } \ \Omega, \tag{4.5}$$

where $1 < p \le 2$ and $I_\delta \in H^1(\Omega)$ such that, as $\delta \to 0$,

$$I_\delta \to I \text{ in } L^2(\Omega), \quad \int_\Omega \alpha|\nabla I_\delta| \to \int_\Omega \alpha|\nabla I|, \tag{4.6}$$

and

$$\|I_\delta\|_{L^\infty(\Omega)} \le \|I\|_{L^\infty}(\Omega), \quad I_\delta|_{\partial\Omega} = I|_{\partial\Omega} = g. \tag{4.7}$$

The existence of I_δ is concluded from Theorem 2.5.

We have the following existence and uniqueness result for the problem (4.3)–(4.5).

LEMMA 4.2. *The problem* (4.3)–(4.5) *admits a unique pseudosolution* $u_{p,\delta} \in L^\infty(0,\infty;W^{1,p}(\Omega)\cap L^2)$, *with* $\partial_t u_{p,\delta} \in L^2(0,\infty;L^2(\Omega))$ *and, for any* $T>0$,

$$\int_0^T\int_\Omega |\partial_t u_{p,\delta}|^2 dxdt + \sup_{t\in[0,T]}\left\{\frac{1}{p}\int_\Omega \alpha|\nabla u_{p,\delta}|^p dx + \frac{1}{2}\int_\Omega |u_{p,\delta}-I_\delta|^2 dx\right\} \le \frac{1}{p}\int_\Omega \alpha|\nabla I_\delta|^p dx. \tag{4.8}$$

This result can be obtained by using the fact that the p-Laplacian operator is a maximal monotone operator.

We can also have the following $L^\infty(\Omega)$ bound for the solution to (4.3)–(4.5).

LEMMA 4.3. *Suppose* $I \in BV(\Omega)\cap L^\infty$, $g \in L^\infty(\partial\Omega)$, *and* $u_{p,\delta}$ *is a weak solution to* (4.3)–(4.5). *Then we have, for any* $T>0$,

$$\|u_{p,\delta}\|_{L^\infty(\Omega^T)} \le \max(\|I\|_{L^\infty(\Omega)}, \|g\|_{L^\infty(\partial\Omega)}). \tag{4.9}$$

Proof. Denote $M = \max(\|I_\delta\|_{L^\infty(\Omega)}, \|g\|_{L^\infty(\partial\Omega)})$. From (4.7),

$$M \le \max(\|I\|_{L^\infty(\Omega)}, \|g\|_{L^\infty(\partial\Omega)}). \tag{4.10}$$

Multiply both sides of (4.3) by $(u_{p,\delta}-M)_+$, where $(u_{p,\delta}-M)_+ = u_{p,\delta}-M$ if $u_{p,\delta}-M\ge 0$ and, otherwise, $(u_{p,\delta}-M)_+ = 0$. Then integrate over Ω to get

$$\int_\Omega \partial_t u_{p,\delta}(u_{p,\delta}-M)_+ dx + \int_{\{u_{p,\delta}\ge M\}} \alpha|\nabla u_{p,\delta}|^p dx$$

$$+\int_\Omega (u_{p,\delta}-M)_+(u_{p,\delta}-I_\delta)dx = 0.$$

Since the last two integrals are nonnegative, we have

$$\int_\Omega \partial_t u_{p,\delta}(u_{p,\delta}-M)_+ dx \le 0. \tag{4.11}$$

Let

$$I(t) = \frac{1}{2}\int_\Omega |(u_{p,\delta}-M)_+|^2 dx.$$

Then $I(0)=0$, and from (4.11)

$$I'(t) = \int_\Omega (u_{p,\delta}-M)_+(\partial_t u_{p,\delta})dx \le 0.$$

Hence $I(t)\le 0$ for all $t\ge 0$, which implies

$$u_{p,\delta}(t) \le M,\ \mathcal{L} \text{ a.e. on } \Omega, \text{ and } \forall t>0.$$

We obtain $u_{p,\delta}(t)\le M$. Similarly, multiplying (4.3) by $(-M-u_{p,\delta})_+$, we can have $u_{p,\delta}(t)\ge -M$. Thus $\|u_{p,\delta}\|_{L^\infty(\Omega^T)}\le M$ for any $T\ge 0$. Then (4.9) follows from (4.10). ∎

Next we shall prove the main theorem regarding the existence, uniqueness, and large time behavior for the solution to (1.3)–(1.5).

THEOREM 4.4. *Let* (H.1)–(H.3) *hold. Assume* $g \in L^\infty(\partial\Omega)$ *and* $I \in BV(\Omega)\cap L^\infty$ *with* $I|_{\partial\Omega} = g$. *Then there exists a unique pseudosolution* $u \in L^\infty(0,\infty; BV(\Omega)\cap L^\infty)$ *to* (1.3)–(1.5). *Moreover, as* $t \to \infty$, *the functions* $u(\cdot,t)$ *converge strongly in* $L^2(\Omega)$ *to a function* $\tilde{u}$, *which minimizes* E^g; *i.e.,* $\tilde{u}$ *is a pseudosolution of* (1.1).

Proof. Let $u_{p,\delta}$ be the weak solution of (4.3)–(4.5). From (4.8)–(4.9), we know that, for fixed $\delta > 0$,

$$u_{p,\delta} \text{ is uniformly bounded in } L^\infty(0,\infty; W^{1,p}(\Omega)\cap L^\infty), \tag{4.12}$$

$$\partial_t u_{p,\delta} \text{ is uniformly bounded in } L^2(\Omega^\infty). \tag{4.12$'$}$$

Now we claim that there exist a sequence of functions $u_{p_j,\delta}$ and a function $u_\delta \in L^\infty(0,\infty; BV(\Omega)\cap L^\infty)$ such that, as $j \to \infty$, $p_j \to 1$,

$$\partial_t u_{p_j,\delta} \to \partial_t u_\delta \text{ weakly in } L^2(\Omega^\infty), \tag{4.13}$$

$$u_{p_j,\delta} \to u_\delta \text{ weakly}^* \text{ in } L^\infty(\Omega^\infty), \tag{4.14}$$

and

$$u_{p_j,\delta} \to u_\delta \text{ in } L^2(\Omega) \text{ uniformly in } t. \tag{4.15}$$

In fact, from (4.12)–(4.12$'$), there is a sequence $u_{p_j,\delta}$ and a function $u_\delta \in L^\infty(\Omega^\infty)$ with $\partial_t u_\delta \in L^2(\Omega^\infty)$ such that (4.13) and (4.14) hold. To see (4.15), note that, for any $f \in L^2(\Omega)$, as $j \to \infty$,

$$\int_\Omega (u_{p_j,\delta}(x,t) - I_\delta(x))f(x)dx = \int_{\Omega^\infty} \partial_s u_{p_j,\delta}(x,s)1_{[0,t]}(s)f(x)dxds$$

$$\to \int_{\Omega^\infty} \partial_s u_\delta(x,s)1_{[0,t]}(s)f(x)dxds = \int_\Omega (u_\delta(x,t) - I_\delta(x))f(x)dx,$$

where $1_{[0,t]}$ is the characteristic function of the set $[0,t] \subset [0,\infty)$. This shows that, for each t,

$$u_{p_j,\delta} \to u_\delta \text{ weakly in } L^2(\Omega). \tag{4.15$'$}$$

By (4.8), for each $t \in [0,\infty)$, $u_{p_j,\delta}(\cdot,t)$ is a bounded sequence in $W^{1,1}(\Omega)$. Hence there is a subsequence of $u_{p_j,\delta}(\cdot,t)$ converging a.e. in Ω and strongly in $L^1(\Omega)$ to u_δ (here we used (4.15$'$)). Since every convergent subsequence of $u_{p_j,\delta}(\cdot,t)$ converges to the same limit, we get that, for each t as $p_j \to 1$,

$$u_{p_j,\delta}(\cdot,t) \to u_\delta(\cdot,t) \text{ in } L^1(\Omega). \tag{4.15$''$}$$

Combining this with (4.9), we get for each t

$$u_{p_j,\delta}(\cdot,t) \to u_\delta(\cdot,t) \text{ in } L^2(\Omega).$$

Furthermore, $t \to u_{p_j,\delta}(\cdot,t) \in L^2(\Omega)$ is equicontinuous since

$$\|u_{p_j,\delta}(\cdot,t) - u_{p_j,\delta}(\cdot,t')\|^2_{L^2(\Omega)} \le |t-t'| \int_{\Omega^\infty} (\partial_t u_{p_j,\delta})^2 dxdt.$$

Then, by a standard argument, we can have the convergence of $u_{p_j,\delta}(\cdot,t)$ to $u_\delta(\cdot,t)$ in $L^2(\Omega)$ uniform in t; this is (4.15). The claim is proved.

Now from (4.8) and (4.15″) we have that

$$\{u_\delta(\cdot,t),\ t\in[0,\infty),\ 1<p\le 2\} \text{ is a bounded set in } BV(\Omega).$$

Moreover, from (4.14),

$$u_\delta \in\ L^\infty(0,\infty;BV(\Omega)\cap L^\infty).$$

Next we show that, for all $v\in L^\infty(0,\infty;H^1(\Omega))$ and each $s\in[0,\infty)$,

$$\int_0^s\int_\Omega \partial_t u_\delta(v-u_\delta)dxdt+\int_0^s\int_\Omega \alpha|\nabla v|dt+1/2\int_0^s\int_\Omega |v-I_\delta|^2dxdt+\int_0^s\int_{\partial\Omega}\alpha|v-g|dH^{n-1}dt$$

$$\ge \int_0^s\int_\Omega \alpha|\nabla u_\delta|dt+1/2\int_0^s\int_\Omega |u_\delta-I_\delta|^2dxdt+\int_0^s\int_{\partial\Omega}|u_\delta-g|dH^{n-1}dt. \tag{4.16}$$

To show this, we first assume further that $v=g$ on $\partial\Omega^\infty$. Multiply (4.3) by $v-u_{p_j,\delta}$, and integrate over Ω^s:

$$\int_0^s\int_\Omega \partial_t u_{p,\delta}(v-u_{p,\delta})dxdt+\int_0^s\int_\Omega \alpha|\nabla u_{p,\delta}|^{p-2}\nabla u_{p,\delta}\cdot\nabla(v-u_{p,\delta})dxdt$$

$$+\int_0^s\int_\Omega (u_{p,\delta}-I_\delta)(v-u_{p,\delta})dxdt=0. \tag{4.17}$$

Write $v-u_{p,\delta}=(v-I_\delta)-(u_{p,\delta}-I_\delta)$, and use the convexity of $|\cdot|^p$ $(p\ge 1)$ to get from (4.17)

$$\int_0^s\int_\Omega \partial_t u_{p,\delta}(v-u_{p,\delta})dxdt+1/p\int_0^s\int_\Omega\alpha|\nabla v|^pdxdt+1/2\int_0^s\int_\Omega|v-I_\delta|^2dxdt$$

$$+\int_0^s\int_{\partial\Omega}|v-g|dH^{n-1}dt\ge 1/p\int_0^s\int_\Omega\alpha|\nabla u_{p,\delta}|^pdxdt+1/2\int_0^s\int_\Omega|u_{p,\delta}-I_\delta|^2dxdt. \tag{4.18}$$

Notice that, using Theorem 2.3, we have

$$\int_\Omega\alpha|\nabla u_\delta|+\int_{\partial\Omega}\alpha|u_\delta-g|dH^{n-1}\le\liminf_{j\to\infty}\int_\Omega\alpha|\nabla u_{p,\delta}|+\int_{\partial\Omega}\alpha|u_{p,\delta}-g|dH^{n-1}$$

$$\le\liminf_{j\to\infty}\left(\int_\Omega\alpha|\nabla u_{p,\delta}|^p\right)^{1/p}\left(\int_\Omega\alpha\right)^{1-\frac{1}{p}}+\int_{\partial\Omega}\alpha|u_{p,\delta}-g|dH^{n-1}$$

$$=\liminf_{j\to\infty}\frac{1}{p}\left(\int_\Omega\alpha|\nabla u_{p,\delta}|^p\right)^{1/p}+\int_{\partial\Omega}\alpha|u_{p,\delta}-g|dH^{n-1}. \tag{4.19}$$

Letting p tend to 1 along p_j in (4.18), using (4.13), (4.15), and (4.19), we get (4.16) for $v\in L^\infty(0,\infty;H^1(\Omega))$ with $v=g$ on $\partial\Omega^\infty$.

To get (4.16) for all $v \in L^\infty(0,\infty;H^1(\Omega))$ (i.e., the condition that $v = g$ on $\partial\Omega^\infty$ is not necessarily satisfied), we let for $\beta > 0$

$$d_\beta(x) = \min(d(x)/\beta, 1),$$

where $d(x)$ is the distance of the point x to the boundary of Ω. Since $|d(x) - d(y)| \le |x - y|$, $d \in W^{1,\infty}$, with $|\nabla d| = 1$. It follows that

$$|\nabla d_\beta| = 1/\beta \text{ if } d(x) < \beta, \text{ and } |\nabla d_\beta| = 0 \text{ if } d(x) \ge \beta.$$

Let

$$v_\beta = d_\beta v + (1 - d_\beta)G \quad \text{for } (x,t) \in \Omega^T.$$

Then

$$v_\beta \in L^\infty(0,\infty;H^1(\Omega)), \text{ and } v_\beta = g \text{ on } \partial\Omega^\infty. \tag{4.20}$$

Therefore, (4.16) holds with v replaced by v_β. It is clear that

$$v_\beta \to v \text{ in } L^2(\Omega^\infty). \tag{4.21}$$

Moreover,

$$\nabla v_\beta = \nabla v_\beta (v - G) + d_\beta \nabla v + (1 - d_\beta)\nabla G.$$

By using Theorem A.1 in the appendix and the fact that $d_\beta \to 1$ as $\beta \to 0$, we get

$$\lim_{\beta\to 0} \int_\Omega \alpha |\nabla v_\beta| = \int_\Omega \alpha |\nabla v| + \int_{\partial\Omega} \alpha |v - g| dH^{n-1}. \tag{4.22}$$

Replacing v by v_β in (4.16), letting $\beta \to 0$, and using (4.20)–(4.22) we get that (4.16) holds for all $v \in L^\infty(0,\infty;H^1(\Omega))$ and each $s \in [0,T]$. Furthermore, by Theorem 2.4, (4.16) also holds for all $v \in L^2(0,\infty;BV(\Omega)\cap L^2)$.

Moreover, replacing u_p by u_{p_j} in (4.8), letting $j \to \infty$ ($p_j \to 1$), and using (4.13), (4.15), (4.19), (4.4), and (4.6), we get

$$\int_0^\infty \int_\Omega |\partial_t u_\delta|^2 dx dt + \sup_{t\in[0,\infty)} \left\{ \int_\Omega \alpha |\nabla u_\delta| + \int_{\partial\Omega} \alpha |u_\delta - g| dH^{n-1} + \frac{1}{2}\int_\Omega |u_\delta - I_\delta|^2 dx \right\}$$

$$\le \int_\Omega \alpha |\nabla I|. \tag{4.23}$$

Next we shall pass to the limit as $\delta \to 0$. Recall from (4.9) and (4.23) that

$$u_\delta \text{ is uniformly bounded in } L^\infty(0,\infty;BV(\Omega)\cap L^\infty), \tag{4.24}$$

$$\partial_t u_\delta \text{ is uniformly bounded in } L^2(\Omega^\infty). \tag{4.25}$$

Then, by an argument similar to the one for getting (4.13)–(4.15), we can find a sequence of functions u_{δ_j} and a function $u \in L^\infty(0,\infty;BV(\Omega)\cap L^\infty)$ such that, as $j \to \infty$, $\delta_j \to 0$,

$$\partial_t u_{\delta_j} \to \partial_t u \text{ weakly in } L^2(\Omega^\infty), \tag{4.26}$$

$$u_{\delta_j} \to u \text{ weakly}^* \text{ in } L^\infty(\Omega^\infty), \tag{4.27}$$

and

$$u_{\delta_j} \to u \text{ in } L^2(\Omega) \text{ uniformly in } t \in [0,\infty). \tag{4.28}$$

Replacing u_δ by u_{δ_j}, letting $j \to \infty$ in (4.16), and using Theorem 2.3, we can have from (4.6), (4.26), and (4.28) that, for all $v \in L^2(0,\infty; BV(\Omega) \cap L^2)$ and each $s \in [0,\infty)$,

$$\int_0^s \int_\Omega \partial_t u(v-u)dxdt + \int_0^s \int_\Omega \alpha|\nabla v|dt + 1/2 \int_0^s \int_\Omega |v-I|^2 dxdt + \int_0^s \int_{\partial\Omega} \alpha|v-g|dH^{n-1}dt$$

$$\geq \int_0^s \int_\Omega \alpha|\nabla u|dt + 1/2 \int_0^s \int_\Omega |u-I|^2 dxdt + \int_0^s \int_{\partial\Omega} |u-g|dH^{n-1}dt. \tag{4.29}$$

We proved the existence of a pseudosolution of (1.3)–(1.5) (see Definition 4.1).

Furthermore, replacing u_δ by u_{δ_j} and letting $j \to \infty$ in (4.23), we get by using (4.26), (4.28), and Theorem 2.3 that the solution u obtained above satisfies the following estimate:

$$\int_0^\infty \int_\Omega |\partial_t u|^2 dxdt + \sup_{t\in[0,\infty)} \left\{ \int_\Omega \alpha|\nabla u| + \int_{\partial\Omega} \alpha|u-g|dH^{n-1} + \frac{1}{2}\int_\Omega |u-I|^2 dx \right\}$$

$$\leq \int_\Omega \alpha|\nabla I|. \tag{4.30}$$

The uniqueness result for (1.3)–(1.5) follows as in [10] and [26]: If u_1 and u_2 are two solutions, one writes the definition of the pseudosolution using each as the function v in (4.2). Adding the resulting inequalities, one finds

$$\int_0^s \int_\Omega \partial_t (u_1 - u_2)^2 \leq 0$$

for all $s > 0$.

At last, we shall show the asymptotic limit of the solution $u(\cdot,t)$ as $t \to \infty$.

Take a function $v \in BV(\Omega) \cap L^2$ in (4.29):

$$\int_\Omega (u(x,s) - u(x,0))v(x)dx - 1/2 \int_\Omega (u^2(x,s) - u^2(x,0))dx + s\int_\Omega \alpha|\nabla v|$$

$$+ s/2 \int_\Omega |v-I|^2 dx + s\int_{\partial\Omega} \alpha|v-g|dH^{n-1} \geq \int_0^s \int_\Omega \alpha|\nabla u|dxdt + 1/2 \int_0^s \int_\Omega |u-I|^2 dxdt$$

$$+ \int_0^s \int_{\partial\Omega} |u-g|dH^{n-1}dt. \tag{4.31}$$

Let

$$w(x,s) = \frac{1}{s}\int_0^s u(x,t)dt.$$

Then, from (4.27) and (4.30), for each s, $w(\cdot,s) \in BV(\Omega) \cap L^\infty$ with uniformly bounded BV and L^∞-norms. Thus there is a sequence $w(\cdot,s_i)$ converging strongly in

$L^1(\Omega)$ and weakly in $BV(\Omega)$ and $L^\infty(\Omega)$ to a function $\hat{w} \in BV(\Omega) \cap L^\infty$ as $s_i \to \infty$. In fact, since $w(\cdot, s_i)$ have uniformly bounded L^∞-norms, the convergence of $w(\cdot, s_i)$ to $\hat{w}$ is strong in $L^2(\Omega)$.

By dividing s in (4.31) and then taking the limit along $s_i \to \infty$, we get that, for any $v \in BV(\Omega) \cap L^2$,

$$\int_\Omega \alpha|\nabla v| + 1/2 \int_\Omega |v - I|^2 dx + \int_{\partial\Omega} \alpha|v - g| dH^{n-1}$$

$$\geq \int_\Omega \alpha|\nabla \hat{w}| + 1/2 \int_\Omega |\hat{w} - I|^2 dx + \int_{\partial\Omega} \alpha|\hat{w} - g| dH^{n-1}.$$

This shows that $\hat{w}$ is the pseudosolution of (1.1). □

Appendix. For $\beta > 0$, let

$$d_\beta(x) = \min(d(x)/\beta, 1),$$

where $d(x)$ is the distance of the point x to the boundary of Ω.

THEOREM A.1. *For each $v \in BV(\Omega) \cap L^\infty$, the vector measures $v\nabla d_\beta$ converge weakly to $-v\gamma dH^{n-1}$ as $\beta \to 0$, where γ is the outward normal to $\partial\Omega$, and*

$$\lim_{\beta\to 0} \int_\Omega |v||\nabla d_\beta| = \int_{\partial\Omega} |v| dH^{n-1}. \tag{A.1}$$

Proof. Without loss of generality, assume $\|v\|_{L^\infty} \leq 1$. By Theorem 3.2.39 in [6], we see that

$$\int_\Omega |v||\nabla d_\beta| \leq \int_\Omega |\nabla d_\beta|$$

is bounded, and as $\beta \to 0$, $\int_\Omega |\nabla d_\beta|$ tends to $H^{n-1}(\partial\Omega)$. Therefore, the family of the vector measures $v\nabla d_\beta$ is uniformly bounded in β as $\beta \to 0$. Let

$$w_\beta = (1 - d_\beta)v.$$

Then $w_\beta \in BV(\Omega)$, and, for each $\phi \in C^1(R^n, R^n)$, by the divergence theorem,

$$\int_\Omega w_\beta \mathrm{div}\phi = -\int_\Omega \phi\nabla w_\beta + \int_{\partial\Omega} (\phi \cdot \gamma) v dH^{n-1}. \tag{A.2}$$

Since $|w_\beta| \leq |v|$ and $w_\beta \to 0$ as $\beta \to 0$, the LHS of (A.2) tends to zero as $\beta \to 0$. Therefore, from (A.2),

$$\lim_{\beta\to 0} \int_\Omega \phi\nabla w_\beta = \int_{\partial\Omega} (\phi \cdot \gamma) v dH^{n-1}. \tag{A.3}$$

Furthermore,

$$\nabla w_\beta = -v\nabla d_\beta + (1 - d_\beta)\nabla v.$$

Since $(1-d_\beta)|\nabla v|$ tends to zero as $\beta \to 0$, and the vector measures $v\nabla d_\beta$ are uniformly bounded in β, we conclude from (A.3) that $v\nabla d_\beta$ converges weakly to $-v\gamma dH^{n-1}$ as $\beta \to 0$.

Using the weak lower semicontinuity of total variation, we get

$$\liminf \int_\Omega |v||\nabla d_\beta| \geq \int_{\partial\Omega} |v| dH^{n-1}. \tag{A.4}$$

Using (A.4) with $1-|v|$ replacing $|v|$, we get

$$\int_{\partial\Omega} dH^{n-1} = \int_{\partial\Omega} |v| dH^{n-1} + \int_{\partial\Omega} (1-|v|) dH^{n-1}$$

$$\leq \liminf \left\{ \int_\Omega |v||\nabla d_\beta| + \int_\Omega (1-|v|)|\nabla d_\beta| \right\}$$

$$\leq \liminf \int_\Omega |\nabla d_\beta| = \int_{\partial\Omega} dH^{n-1}. \tag{A.5}$$

Therefore, each inequality in (A.5) should be an equality. Furthermore, noticing that $\int_{\partial\Omega} |v| dH^{n-1} \leq \liminf \int_\Omega |v||\nabla d_\beta|$ and $\int_{\partial\Omega}(1-|v|)dH^{n-1} \leq \liminf \int_\Omega (1-|v|)|\nabla d_\beta|$, we obtain (A.1). □

Let

$$a_p = \inf_{v \in W^{1,p}(\Omega)\cap L^2} E_p^g(v) \text{ for } p>1, \text{ and } a_1 = \inf_{v\in BV(\Omega)\cap L^2} E^g(v),$$

$$A_p = \inf_{v \in W_g^{1,p}\cap L^2} E_p(v) \text{ for } p>1, \text{ and } A_1 = \inf_{v\in BV_g\cap L^2} E(v),$$

where $E(v)$, $E^g(v)$, $E_p(v)$, and $E_p^g(v)$ are as defined in (1.1)–(1.2) and (2.5)–(2.6), respectively. If $p>1$, both infima a_p and A_p are attained. For $p=1$, a_1 is attained. We will show below that $a_1 = A_1$. This will justify the definition of the pseudosolution for (1.1). We will use the following simple inequality:

For $b>0$ and $1 \leq q \leq p < \infty$,

$$\frac{b^q}{q} \leq \frac{b^p}{p} + \frac{p-q}{pq}. \tag{A.6}$$

This can be proved by using the Hölder inequality

$$b^q \leq \frac{qb^p}{p} + \frac{p-q}{pq}.$$

Proposition A.2. *For $p \geq 1$, both a_p and A_p are right continuous in p. We also have for $1 \leq q \leq p < \infty$*

$$a_q \leq \frac{p-q}{pq}|\Omega| + a_p, \tag{A.7}$$

$$A_q \leq \frac{p-q}{pq}|\Omega| + A_p, \tag{A.7$'$}$$

Proof. (1) Let $1 \leq q \leq p < \infty$ and $v_p \in W^{1,p}(\Omega)\cap L^2$ if $p>1$, and $v_p \in v \in BV(\Omega)\cap L^2$ if $p=1$. Then, using (A.6),

$$a_q \leq E_q^g(v_p) = 1/q \int_\Omega \alpha |\nabla v_p|^q + 1/2 \int_\Omega |v_p - I|^2 dx + \int_{\partial\Omega} \alpha |v_p - g| dH^{n-1}$$

$$\leq \frac{p-q}{pq}|\Omega| + E_p^g(v_p).$$

Since this is valid for all $v_p \in W^{1,p}(\Omega) \cap L^2$ if $p > 1$ and $v_p \in BV(\Omega) \cap L^2$ if $p = 1$, we get (A.7). (A.7′) can be proved by the same argument.

(2) Now we prove the right continuity of a_p, i.e.,

$$\lim_{p \downarrow q} a_p = a_q, \quad q \geq 1.$$

For $\epsilon > 0$, assume $u \in W^{1,\infty}(\Omega) \cap L^2$ such that

$$E_q^g(u) \leq a_q + \epsilon.$$

Then

$$\limsup_{p \downarrow q} a_p \leq \limsup_{p \downarrow q} E_p^g(u) = E_q^g(u) \leq a_q + \epsilon.$$

Let $\epsilon \to 0$, and we get

$$\limsup_{p \downarrow q} a_p \leq a_q. \tag{A.8}$$

On the other hand, by taking liminf as $p \downarrow q$ on both sides of (A.7),

$$a_q \leq \liminf_{p \downarrow q} a_p.$$

Combining this with (A.8) proves the right continuity of a_p for $p \geq 1$.

(3) To prove the right continuity of A_p, we need some modification for the argument applied above for a_p since the inequality similar to (A.9) is not valid for $q = 1$ due to the fact that $W_g^{1,\infty}$ is not dense in BV_g, while $W^{1,\infty}$ is weakly dense in BV. However, in this case, we can use Theorem 2.5 to get a similar result as follows: Let $\epsilon > 0$, and assume that $u \in BV_g \cap L^2$ such that

$$E(u) \leq A_1 + \epsilon. \tag{A.9}$$

By using (A.9) and Theorem 2.5, there is $v \in W_g^{1,2} \cap L^2$ such that

$$E(v) \leq E(u) + \epsilon \leq A_1 + 2\epsilon. \tag{A.10}$$

For $q > 1$, it is clear that there is a function $v \in W_g^{1,\infty} \cap L^2$ such that

$$E_q(v) \leq a_q + \epsilon. \tag{A.11}$$

Now repeating the argument in the second step and using (A.10)–(A.11) give the right continuity of A_p for $p \geq 1$. □

THEOREM A.3. *Let* (H.1)–(H.3) *hold. Then*

$$\inf_{v \in BV(\Omega) \cap L^2} E^g(v) = \inf_{v \in BV_g \cap L^2} E(v). \tag{A.12}$$

Proof. Since $BV_g \subset BV(\Omega)$,

$$\inf_{v \in BV(\Omega) \cap L^2} E^g(v) \leq \inf_{v \in BV_g \cap L^2} E(v).$$

Next we shall prove

$$\inf_{v \in BV(\Omega) \cap L^2} E^g(v) \geq \inf_{v \in BV_g \cap L^2} E(v). \tag{A.13}$$

to get (A.12).

From Theorem 3.2, we know that there exists a unique $u \in BV(\Omega) \cap L^2$ such that $E^g(u) = \inf_{v \in BV(\Omega) \cap L^2} E^g(v)$. Let $d_\beta(x)$ be the function defined in (4.20). Define

$$u_\beta = d_\beta u + (1 - d_\beta)G.$$

Then $u_\beta \in BV_g$ so that

$$\inf_{v \in BV_g \cap L^2} E(v) \leq E(u_\beta). \tag{A.14}$$

Furthermore, noticing that $u_\beta = u$ on $\{d(x) \geq \beta\}$, we have

$$E(u_\beta) \leq \int_{\Omega \cap \{d(x) \geq \beta\}} \alpha |\nabla u| + \int_{\Omega \cap \{d(x) < \beta\}} \alpha |u - G| |\nabla d_\beta|$$

$$+ \int_{\Omega \cap \{d(x) < \beta\}} \alpha |\nabla (u - G)| |d_\beta| + \int_{\Omega \cap \{d(x) < \beta\}} \alpha |\nabla G| + \int_\Omega |u_\beta - I|^2. \tag{A.15}$$

As $\beta \to 0$, $u_\beta \to u$ in $L^2(\Omega)$, and $|\{\Omega \cap \{d(x) \leq \beta\}\}| \to 0$. Using these facts and (A.1) with $v = u - G$, we get from (A.15) that

$$\lim_{\beta \to 0} E(u_\beta) \leq \int_\Omega \alpha |\nabla u| + \int_{\partial \Omega} \alpha |u - g| dH^{n-1} + \int_\Omega |u - I|^2.$$

The combination of this with (A.14) leads to (A.13) and hence to (A.12). □

REFERENCES

[1] R. Acar and C. R. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems, 10 (1994), pp. 1217–1229.

[2] C. Bouman and K. Sauer, *A generalized Gaussian image model for edge-preserving MAP estimation*, IEEE Trans. Image Process., 2 (1993), pp. 296–310.

[3] P. B. Omgren and T. F. Chan, *Color TV: Total variation methods for restoration of vector-valued images*, IEEE Trans. Image Process., 7 (1998), pp. 304–309.

[4] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.

[5] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.

[6] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[7] J. Heinonen, T. Kilpeläinen, and O. Martio, *Nonlinear Potential Theory of Degenerate Elliptic Equations*, Oxford Math. Monogr., Oxford University Press, New York, 1993.

[8] R. Hardt and F. H. Lin, *Mappings minimizing the L^p norm of the gradient*, Comm. Pure Appl. Math., 40 (1987), pp. 555–588.

[9] P. Hartman and G. Stampacchia, *On some non-linear elliptic differential functional equations*, Acta Math., 115 (1966), pp. 271–310.

[10] R. Hardt and X. Zhou, *An evolution problem for linear growth functionals*, Comm. Partial Differential Equations, 19 (1994), pp. 1879–1907.

[11] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Notes on Pure Math. 10, Department of Pure Mathematics, Australian National University, Canberra, Australia, 1977.

[12] H. Parks, *Explicit determination of area minimizing hypersurfaces*, Duke Math. J., 44 (1977), pp. 519–534.

[13] H. Parks, *Explicit determination of area minimizing hypersurfaces* II, Mem. Amer. Math. Soc., 342 (1986).

[14] H. Parks and W. Ziemer, *Jacobi fields and functions of least gradient*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 11 (1984), pp. 505–527.

[15] L. Rudin and S. J. Osher, *Total variation based image restoration with free local constraints*, in Proceedings of the IEEE International Conference on Image Processing, Austin, TX, 1994, pp. 31–35.

[16] L. Rudin, S. J. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.

[17] D. M. Strong, *Adaptive Total Variation Minimizing Image Restoration*, Thesis, University of California, Los Angeles, CA, 1997, CAM97-38(UCLA).

[18] D. M. Strong and T. F. Chan, *Spatially and Scale Adaptive Total Variation Based Regularization and Anisotropic Diffusion in Image Processing*, Technical report, University of California, Los Angeles, CA, CAM96-46(UCLA).

[19] D. M. Strong and T. F. Chan, *Edge-Preserving and Scale-Dependent Properties of Total Variation Regularization*, Technical report, University of California, Los Angeles, CA, CAM00-38(UCLA).

[20] P. Sternberg, G. Williams, and W. P. Ziemer, *Existence, uniqueness and regularity for functions of least gradient*, J. Reine Angew. Math., 430 (1992), pp. 35–60.

[21] P. Sternberg and W. P. Ziemer, *The Dirichlet problem for functions of least gradient*, in Degenerate Diffusions, IMA Vol. Math. Appl. 47, Springer-Verlag, New York, 1993, pp. 197–214.

[22] R. Temam, *Solution généralisées de certain équations du type hypersurfaces minimales*, Arch. Ration. Mech. Anal., 44 (1971), pp. 121–156.

[23] L. A. Vese and S. J. Osher, *Numerical Methods for p-Harmonic Flows and Applications to Image Processing*, Technical report, University of California, Los Angeles, CA, CAM01-22(UCLA).

[24] C. R. Vogel and M. E. Oman, *Iterative methods for total variation denoising*, SIAM J. Sci. Comput., 17 (1996), pp. 227–238.

[25] L. Vese, *A study in the BV space of a denoising-deblurring variational problem*, Appl. Math. Optim., 44 (2001), pp. 131–161.

[26] X. Zhou, *An evolution problem for plastic antiplanar shear*, Appl. Math. Optim., 25 (1992), pp. 263–285.

SIAM J. MATH. ANAL.
Vol. 34, No. 5, pp. 1105–1127

VORTICES IN p-WAVE SUPERCONDUCTIVITY*

FANGHUA LIN† AND TAI-CHIA LIN‡

Abstract. In the theory of p-wave superconductivity, the Ginzburg–Landau energy functionals with multicomponent order parameters were employed. Here we find a minimizer of a reduced form of the p-wave Ginzburg–Landau free energy with two-component order parameters. The minimizer has distinct degree-one (or minus one) vortices in each component. We also derive a system of ordinary differential equations as the motion equations of vortices in the approximated gradient flow for p-wave superconductivity.

Key words. vortices, dynamics, p-wave superconductivity, Ginzburg–Landau

AMS subject classifications. 35J35, 82D55

PII. S0036141001395820

1. Introduction.

It is well known that many of the heavy-fermion superconductors are thought to represent a novel form of superconductivity. Remarkable evidence in support of an unconventional superconducting state in the heavy-fermion superconductors has accumulated from specific heat, upper critical field, and various transport measurements, all of which show anomalous properties compared with those of conventional superconductors (cf. [2], [7], [22], [23], [25]). Conventional superconductors refer to those with the pairing symmetry of the s-wave and the spin singlet. However, it is widely accepted that an anisotropic p-wave spin-triplet pairing may be realized in heavy-fermion superconductors.

The possibility of the p-wave spin-triplet pairing has been investigated since the 1970's for superfluid ^3He, a heavy-fermion system UPt_3, or, most recently, an oxide Sr_2RuO_4 (cf. [12], [19], [28], [29]). The strongest evidence for unconventional superconductivity comes from the multiple superconducting phases of UPt_3. There are two superconducting phases in the zero field (cf. [7]).

To describe p-wave superconductors, we consider a simple situation with two-component order parameters η_i, $i = 1, 2$. In the absence of a magnetic field, the Ginzburg–Landau free energy is given by

$$F(\eta_1,\eta_2)=\int_{\mathbb{R}^2} \begin{aligned} &K_1(|\partial_x\eta_1|^2+|\partial_y\eta_2|^2)+K_2(|\partial_x\eta_2|^2+|\partial_y\eta_1|^2)+f_{pot}(\eta_1,\eta_2)\\ &+K_3\left(\partial_x\eta_1^*\partial_y\eta_2+\text{c.c.}\right)+K_4\left(\partial_x\eta_2^*\partial_y\eta_1+\text{c.c.}\right)dx\,dy, \end{aligned} \tag{1.1}$$

for η_1 and η_2 are complex-valued order parameters, where $K_j, j = 1, \dots, 4$, are material constants and the asterisk denotes the complex conjugate. Hereafter,

$$f_{pot}(\eta_1,\eta_2)=-\alpha_0\left(|\eta_1|^2+|\eta_2|^2\right)+\alpha_1\left(|\eta_1|^2+|\eta_2|^2\right)^2+\alpha_2\left(\eta_1^*\eta_2-\eta_1\eta_2^*\right)^2, \tag{1.2}$$

*Received by the editors September 20, 2001; accepted for publication (in revised form) September 8, 2002; published electronically April 15, 2003.
http://www.siam.org/journals/sima/34-5/39582.html

†Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, 817, New York, NY 10012-0711 (linf@cims.nyu.edu). The work of this author was partially supported by NSF grant DMS 9896391.

‡Department of Mathematics, Chung-Cheng University, Chia-Yi, Taiwan (tclin@math.ccu.edu.tw). The work of this author was partially supported by NSC90-2115-M-194015.

where $\alpha_j, j = 0, 1, 2$, are constants. Zhu et al. [30] derived (1.1) from reasonable microscopic models describing Sr_2RuO_4. The expression of the Ginzburg–Landau free energy (1.1) agrees quite well with that constructed from the group-theoretical argument (cf. [24]) for the Γ_5^- superconducting state in the tetragonal D_{4h} (except one coefficient) and hexagonal symmetry D_{6h}.

In this paper, we assume that $K_1 = K_2 = K > 0$ and $K_3 = K_4 = 0$ to avoid various complications which come from additional derivative terms. Then (1.1) becomes

$$F(\eta_1, \eta_2) = \int_{\mathbb{R}^2} K(|\nabla \eta_1|^2 + |\nabla \eta_2|^2) + f_{pot}(\eta_1, \eta_2)\, dx\, dy. \tag{1.3}$$

Moreover, it is easy to check that

$$f_{pot}(\eta_1, \eta_2) = -\alpha_0 (|\eta_1|^2 + |\eta_2|^2) + \beta_1 (|\eta_1|^2 + |\eta_2|^2)^2 + \beta_2 |\eta_1^2 + \eta_2^2|^2, \tag{1.4}$$

where $\beta_1 = \alpha_1 - \alpha_2, \beta_2 = \alpha_2$. From [11], we use the following convenient parametrization of the order parameter:

$$(\eta_1, \eta_2)(x, y) = f(x, y)\, (N\, \cos\phi + i\, M\, \sin\phi), \tag{1.5}$$

where $f = f(x, y), \phi = \phi(x, y)$ are real-valued functions and $N = N(x, y) = (N_1, N_2)$, $M = M(x, y) = (M_1, M_2)$ are S^1-valued functions. Then (1.4) becomes

$$f_{pot} = -\alpha_0\, f^2 + \beta_1\, f^4 + \beta_2\, f^4\, [\cos^2 2\phi + (M \cdot N)^2 \sin^2 2\phi], \tag{1.6}$$

which can be easily minimized to give the two phases as follows:

(i) *Phase* I. As $\beta_2 > 0, (\eta_1, \eta_2) = f\, \frac{N + i\, M}{\sqrt{2}}\, , N \perp M\, , \phi = \pi/4$.

(ii) *Phase* II. As $\beta_2 < 0, (\eta_1, \eta_2) = f\, e^{i\, \phi}\, N\, , N = M$.

In phase I, we set $M = \pm(-N_2, N_1)$ and $\psi = f\, (N_1 + i\, N_2)$. Then (1.3) becomes

$$F(\psi) = \int_{\mathbb{R}^2} K\, |\nabla\, \psi|^2 - \alpha_0\, |\psi|^2 + \beta_1\, |\psi|^4\, dx\, dy, \tag{1.7}$$

for ψ is a complex-valued function, where K, α_0, and β_1 are positive constants. We may rescale ψ and spatial variables suitably and then transform (1.7) into (up to some constants) the s-wave Ginzburg–Landau free energy (cf. [8]) given by

$$\int_{\mathbb{R}^2} \frac{1}{2}|\nabla\psi|^2 + \frac{1}{4}(1 - |\psi|^2)^2\, .$$

Moreover, we may approximate the s-wave Ginzburg–Landau free energy by

$$\int_{\frac{1}{\epsilon}\Omega} \frac{1}{2}|\nabla\psi|^2 + \frac{1}{4}(1 - |\psi|^2)^2\, ,$$

where $0 < \epsilon \ll 1$ is a small parameter and Ω is a bounded smooth domain in $\mathbb{R}^2$. Then we rescale the spatial variables by ϵ and obtain the energy functional as follows:

$$E_\epsilon(\psi) = \int_\Omega \frac{1}{2}\, |\nabla\, \psi|^2 + \frac{1}{4\, \epsilon^2}\, (1 - |\psi|^2)^2. \tag{1.8}$$

The Euler–Lagrange equation and the gradient flow of (1.8) with the Dirichlet boundary condition have been investigated as the fundamental equations for understanding

s-wave superconductors; see Bethuel, Brezis, and Helein [1], Struwe [26], [27], Lin [14], Lin and Lin [17], Pacard and Riviére [20], and many others.

In phase II, we set $\psi = f\, e^{i\,\phi}$. Then (1.3) becomes

$$F(\psi, N) = \int_{\mathbb{R}^2} K\,(|\nabla\,\psi|^2 + |\psi|^2\,|\nabla N|^2) - \alpha_0\,|\psi|^2 + (\beta_1 + \beta_2)\,|\psi|^4\,dx\,dy, \tag{1.9}$$

for ψ is a complex-valued function and N is an S^1-valued function. Hereafter, we assume that $\beta_1+\beta_2 > 0$ in order to ensure the stability of the phase II superconductivity. As for (1.8), we may, after proper normalization, put (1.9) in the form

$$G_\epsilon(\psi, N) = \int_\Omega \frac{1}{2}\,(|\nabla\,\psi|^2 + |\psi|^2\,|\nabla N|^2) + \frac{1}{4\,\epsilon^2}\,(1 - |\psi|^2)^2\,dx\,dy. \tag{1.10}$$

Note that when ψ is real-valued, (1.10) can be regarded as a model of nematic liquid crystals (cf. [6]). Let $u = |\psi|\,Z \in \mathbb{C}$, where $Z = (N_1 + i\,N_2) \in S^1$ in $\mathbb{C}$ and $N = (N_1, N_2) \in S^1$ in $\mathbb{R}^2$. Then we may rewrite (1.10) as

$$E_\epsilon(\psi, u) = \int_\Omega \frac{1}{2}\,(|\nabla\,\psi|^2 + |\nabla\,u|^2 - |\nabla\,|\psi||^2) + \frac{1}{4\,\epsilon^2}\,(1 - |\psi|^2)^2\,, \tag{1.11}$$

for ψ and u are complex-valued functions satisfying $|\psi| = |u|$.

Vortex configurations and vortex dynamics in superconductivity are physically meaningful problems. One way to create vortices of the energy minimizers and the solutions of gradient flows without external fields is to impose the Dirichlet boundary condition of ψ and u; see also [15] for the Neumann boundary condition. In this paper, we study the asymptotic behavior of the energy minimizer of (1.11) with a given Dirichlet boundary condition as $\epsilon \to 0+$. Using the idea introduced by the first author in studying the dynamics of Ginzburg–Landau vortices (see, e.g., [14]), we also derive the corresponding dynamical law of vortices for p-wave superconductivity with two-component order parameters.

Let $(\psi_\epsilon, u_\epsilon)$ be the energy minimizer of (1.11) for $\psi \in H^1_\eta(\Omega;\mathbb{C})$, $u \in H^1_g(\Omega;\mathbb{C})$, and $|\psi| = |u|$ in Ω. Hereafter, $H^1_\eta(\Omega;\mathbb{C}) = \{v \in H^1(\Omega;\mathbb{C}) : v = \eta \text{ on } \partial\Omega\}$, and $H^1_g(\Omega;\mathbb{C}) = \{w \in H^1(\Omega;\mathbb{C}) : w = g \text{ on } \partial\Omega\}$, where η and g are given smooth maps from $\partial\Omega$ into S^1 with degrees d_1 and d_2, respectively. We shall assume, by simply taking the complex conjugate if it is needed, that $d_1, d_2 \geq 0$. The case in which either d_1 or d_2 vanishes will be seen to be very easy. Essentially, all the statements in this case follow from earlier works (cf. [1], [14]). In the case that both d_1 and d_2 are positive, the general picture will be as follows: ψ_ϵ has d_1 degree-one vortices and u_ϵ has d_2 degree-one vortices in Ω. We shall denote essential zeros (see Proposition 2.2) as degree-one vortices. Moreover, there is an integer $0 \leq N_0 \leq \min(d_1, d_2)$, N_0 distinct points $a_1, \ldots, a_{N_0} \in \Omega$, $d_1 - N_0$ distinct points $b_1, \ldots, b_{d_1-N_0} \in \Omega$, and $d_2 - N_0$ distinct points $c_1, \ldots, c_{d_2-N_0} \in \Omega$ such that the essential zeros of ψ_ϵ converge (as $\epsilon \to 0+$) to $a_1, \ldots, a_{N_0}, b_1, \ldots, b_{d_1-N_0}$, and the essential zeros of u_ϵ converges (as $\epsilon \to 0+$) to $a_1, \ldots, a_{N_0}, c_1, \ldots, c_{d_2-N_0}$. Due to the constraint $|\psi_\epsilon| = |u_\epsilon|$, ψ_ϵ has nonessential (i.e., degree-zero) zeros near $c_1, \ldots, c_{d_2-N_0}$, and u_ϵ has nonessential zeros near $b_1, \ldots, b_{d_1-N_0}$. The number N_0 may depend on the domain Ω and the boundary conditions η and g. The reader may find various situations in which the number N_0 may be zero or may be equal to d_1 or d_2. Some such cases may be found in Remark 3.1. The dynamical law that governs motions of vortices will be a system of ordinary differential equations for those corresponding point vortices $a_1(t), \ldots, a_{N_0}(t), b_1(t), \ldots, b_{d_1-N_0}(t), c_1(t), \ldots, c_{d_2-N_0}(t)$ for time $t > 0$. In our forthcoming paper, we shall address the multicomponent p-wave superconductivity.

2. Basic energy estimates. We start with the following proposition (see [16]).

PROPOSITION 2.1. *Suppose* $v_\epsilon \in H^1_z(\Omega;\mathbb{C}) \equiv \{v \in H^1(\Omega;\mathbb{C}) : v = z \text{ on } \partial\Omega\}$ *such that*

$$E_\epsilon(v_\epsilon) \le \pi\, d \log \frac{1}{\epsilon} + C_0 \qquad \text{as } \epsilon \to 0+, \tag{2.1}$$

where C_0 *is a positive constant independent of* ϵ *and* $z : \partial\Omega \to S^1$ *is a smooth map with degree* $d > 0$. *Here*

$$E_\epsilon(v_\epsilon) = \int_\Omega e_\epsilon(v_\epsilon)\,,$$
$$e_\epsilon(v_\epsilon) = \frac{1}{2}\left[|\nabla v_\epsilon|^2 + \frac{1}{2\,\epsilon^2}\,(1-|v_\epsilon|^2)^2\right].$$

Then there are exactly d *distinct points* $a^\epsilon_j \in \Omega, j = 1,\dots,d$, *of* v_ϵ *such that* $a^\epsilon_j \to a_j \in \Omega$ *(up to a subsequence) as* $\epsilon \to 0+$,

$$\epsilon^{\alpha_j} \int_{\partial B_j} e_\epsilon(v_\epsilon) \le C_1 \qquad \text{and} \qquad \deg\left(\frac{v_\epsilon}{|v_\epsilon|}, \partial B_j\right) = 1, \tag{2.2}$$

where $B_j = B_{\epsilon^{\alpha_j}}(a^\epsilon_j)$ *for* $j = 1,\dots,d$, $0 < \alpha_j < 1$, *and* $C_1 > 0$ *is a universal constant. Furthermore,*

$$\min\{|a_i - a_j|,\, dist(a_i, \partial\Omega) : i, j = 1,\dots,d, i \ne j\} \ge \delta_0(z,\Omega,C_0) > 0,$$

and v_ϵ *converges (up to a subsequence) to a map of the form*

$$\prod_{j=1}^{d} \frac{x - a_j}{|x - a_j|}\, e^{i\,h(x)}$$

strongly in $L^2(\Omega)$ *and weakly in* $H^1_{loc}(\bar\Omega\backslash\{a_1,\dots,a_d\})$ *as* $\epsilon \to 0+$. *Moreover,*

$$\|h\|_{H^1(\Omega)} \le C(C_0, z, \Omega).$$

We shall call these a^ϵ_j's essential zeros of v_ϵ. It is obvious that these essential zeros are well defined (up to a possible error of ϵ to a fixed positive power; see [14]).

To study (1.11), we may decompose it as

$$E_\epsilon(\psi, u) = \tilde E_\epsilon(\psi) + \tilde E_\epsilon(u) + \frac{1}{6}\int_\Omega |\nabla\,|u||^2, \tag{2.3}$$

where

$$\tilde E_\epsilon(u) = \int_\Omega \frac{1}{2}\left(|\nabla u|^2 - \frac{2}{3}\,|\nabla\,|u||^2\right) + \frac{1}{8\,\epsilon^2}\,(1-|u|^2)^2, \tag{2.4}$$

for ψ and u are complex-valued functions with $|\psi| = |u|$.

The following lower bound for $\tilde E_\epsilon$ is crucial as in [1].

PROPOSITION 2.2. *Suppose* v_ϵ *is a minimizer of* $\tilde E_\epsilon$ *over* $H^1_z(\Omega;\mathbb{C}) \equiv \{v \in H^1(\Omega;\mathbb{C}) : v = z \text{ on } \partial\Omega\}$, *where* $z : \partial\Omega \to S^1$ *is a smooth map with degree* $d > 0$. *Then* v_ϵ *has exactly* d *essential zeros* $a^\epsilon_j, j = 1,\dots,d$, *in* Ω; *i.e., there are exactly* d *balls, say,* $B_j = B_{\epsilon^{\alpha_j}}(a^\epsilon_j)$, $\alpha_j \in (0,1), j = 1,\dots,d$, *such that* $\deg(\frac{v_\epsilon}{|v_\epsilon|}, \partial B_j) = 1$

and $\epsilon^{\alpha_j} \int_{\partial B_j} \tilde{e}_\epsilon(v_\epsilon) \le K_j$ *for* $j = 1, \dots, d$, *where* $K_j = K_j(\alpha_j) > 0$ *are constants. Hereafter,* $\tilde{e}_\epsilon$ *is the energy density of* $\tilde{E}_\epsilon$ *defined by*

$$\tilde{e}_\epsilon(u) = \frac{1}{2}\left(|\nabla u|^2 - \frac{2}{3}|\nabla |u||^2\right) + \frac{1}{8\,\epsilon^2}(1-|u|^2)^2\,. \tag{2.5}$$

Moreover,

$$\tilde{E}_\epsilon(v_\epsilon) = \pi\, d\, \log\frac{1}{\epsilon} + O(1) \quad \text{as } \epsilon \to 0+\,, \tag{2.6}$$

where $O(1)$ *is independent of* ϵ.

From (2.3), Proposition 2.2, and Lecture 1 of [14], we have the following corollary.

COROLLARY 2.3. *Let* $\psi_\epsilon \in H^1_\eta(\Omega;\mathbb{C})$ *and* $u_\epsilon \in H^1_g(\Omega;\mathbb{C})$. *Then*

$$E_\epsilon(\psi_\epsilon, u_\epsilon) \le \pi\,(d_1+d_2)\,\log\frac{1}{\epsilon} + O(1) \tag{2.7}$$

if and only if

$$E_\epsilon(\psi_\epsilon) \le \pi\, d_1\, \log\frac{1}{\epsilon} + O(1), \quad E_\epsilon(u_\epsilon) \le \pi\, d_2\, \log\frac{1}{\epsilon} + O(1).$$

By Proposition 2.2, (2.3), and (2.7), it is also easy to check that

$$\int_\Omega |\nabla\,|\psi_\epsilon||^2 = \int_\Omega |\nabla\,|u_\epsilon||^2 \le M_5, \tag{2.8}$$

where M_5 is a positive constant independent of ϵ. Moreover, by (2.7), (2.8), and Lecture 1 of [14], we can write

$$E_\epsilon(\psi_\epsilon) \le \pi\, d_1\, \log\frac{1}{\epsilon} + M_6\,, \quad E_\epsilon(u_\epsilon) \le \pi\, d_2\, \log\frac{1}{\epsilon} + M_6, \tag{2.9}$$

where M_6 is a positive constant independent of ϵ.

The proof of Proposition 2.2 is based on Lemma 2.2 of [16], Theorem 3.1, and the structure theorem of [14]. One may find another proof using techniques of Theorems 2 and 3 in [21]. For the sake of completeness, we give the proof of Proposition 2.2.

Proof of Proposition 2.2. From [13], $\tilde{E}_\epsilon$ has a minimizer v_ϵ over the space $H^1_z(\Omega;\mathbb{C})$. Moreover, v_ϵ is Lipschitz continuous on Ω. Let U_ϵ be the minimizer of E_ϵ (defined in (1.8)) over $H^1_z(\Omega;\mathbb{C})$. Then it is well known (see [1]) that $E_\epsilon(U_\epsilon) \le \pi\, d \log\frac{1}{\epsilon} + M_0$, where M_0 is a positive constant. Hence it is obvious that

$$\tilde{E}_\epsilon(v_\epsilon) \le \tilde{E}_\epsilon(U_\epsilon) \le E_\epsilon(U_\epsilon) \le \pi\, d\, \log\frac{1}{\epsilon} + M_0. \tag{2.10}$$

By the energy comparison, it is easy to show that

$$|v_\epsilon| \le 1 \quad \text{in } \Omega. \tag{2.11}$$

To obtain d essential zeros of v_ϵ, we need the following lemma.

LEMMA 2.4. *Suppose* $|v_\epsilon(a_1)| < \frac{1}{2}$, *where* $a_1 \in \Omega$. *Then there exists* $\alpha_1 \in (0,1)$ *independent of* $\epsilon \le \epsilon_o$ *for some small but fixed positive* ϵ_o *such that* $\deg(\frac{v_\epsilon}{|v_\epsilon|}, \partial B_1) \ne 0$ *and* $\epsilon^{\alpha_1} \int_{\partial B_1} \tilde{e}_\epsilon(v_\epsilon) \le K_1$, *where* $B_1 = B_{\epsilon^{\alpha_1}}(a_1)$ *and* $K_1 = K_1(\alpha_1) > 0$ *is a universal constant.*

Assume Lemma 2.4 for the moment; we continue the proof of Proposition 2.2 as follows. By (2.10), (2.11), Lemma 2.4, and the proof of the structure theorem in [14], v_ϵ has only d essential zeros a_j^ϵ's in Ω, and $\deg(\frac{v_\epsilon}{|v_\epsilon|}, \partial B_j) = 1$ for $j = 1, \dots, d$, where $B_j = B_{\epsilon^{\alpha_j}}(a_j^\epsilon)$ and $\alpha_j \in (0,1)$. Moreover, by the same argument of Lemma 2.2 in [16], we obtain

$$\int_{\Omega \setminus \cup_{j=1}^d B_j} \tilde{e}_\epsilon(v_\epsilon) \geq \pi \sum_{j=1}^d \alpha_j \log \frac{1}{\epsilon} - M_1, \tag{2.12}$$

and hence

$$\int_{B_j} \tilde{e}_\epsilon(v_\epsilon) \leq \pi (1-\alpha_j) \log \frac{1}{\epsilon} + M_1 \quad \text{for } j = 1, \dots, d, \tag{2.13}$$

where M_1 is a positive constant independent of ϵ.

Now we claim that

$$\int_{B_j} \tilde{e}_\epsilon(v_\epsilon) \geq \pi (1-\alpha_j) \log \frac{1}{\epsilon} - K \quad \text{for } j = 1, \dots, d, \tag{2.14}$$

where K is a positive constant independent of ϵ. Without loss of generality, we may assume that $B_j = B_{\theta_0}(0), \theta_0 = \epsilon^{\alpha_j}$. Then (2.13) implies that

$$\int_{B_{\theta_0}(0)} \tilde{e}_\epsilon(v_\epsilon) \leq \pi \log \frac{\theta_0}{\epsilon} + M_1. \tag{2.15}$$

Moreover, we may rescale the spatial variable and rewrite (2.13) as

$$\int_{B_1(0)} \tilde{e}_{\epsilon_1}(v_\epsilon) \leq \pi (1-\alpha_j) \log \frac{1}{\epsilon} + M_1, \tag{2.16}$$

where $\epsilon_1 = \epsilon^{1-\alpha_j}$. By (2.16) and the Fubini theorem (cf. [15]), there exists $\theta_1 \in (\epsilon^{2\alpha_j}, \epsilon^{\alpha_j})$ such that

$$\theta_0 \theta_1 \int_{\partial B_{\theta_0 \theta_1}(0)} \tilde{e}_\epsilon (v_\epsilon) \leq C(\alpha_j, M_1)$$

and that

$$\deg\left(\frac{v_\epsilon}{|v_\epsilon|}, \partial B_{\theta_0 \theta_1}(0)\right) = 1.$$

Hence, by the same argument of Lemma 2.2 in [16], we have

$$\int_{B_{\theta_0}(0) \setminus B_{\theta_0\, \theta_1}(0)} \tilde{e}_\epsilon(v_\epsilon) \geq \pi \log \frac{1}{\theta_1} - M_2 \tag{2.17}$$

and

$$\int_{B_{\theta_0\, \theta_1}(0)} \tilde{e}_\epsilon(v_\epsilon) \leq \pi \log \frac{\theta_0\, \theta_1}{\epsilon} + M_1 + M_2, \tag{2.18}$$

where M_2 is a positive constant satisfying $M_2 \le C_0\,\theta_0$. Here C_0 is a positive constant independent of ϵ. Thus, by induction, we may obtain $\theta_1, \dots, \theta_m \in (\epsilon^{2\alpha_j}, \epsilon^{\alpha_j})$ such that $\epsilon = \theta_0\theta_1\cdots\theta_m$ and

$$\int_{B_{\theta_0\cdots\theta_{k-1}}(0)\setminus B_{\theta_0\cdots\theta_k}(0)} \tilde{e}_\epsilon(v_\epsilon) \ge \pi \log\frac{1}{\theta_k} - M_{k+1} \tag{2.19}$$

and

$$\int_{B_{\theta_0\cdots\theta_k}(0)} \tilde{e}_\epsilon(v_\epsilon) \le \pi \log\frac{\theta_0\cdots\theta_k}{\epsilon} + \sum_{j=1}^{k+1} M_j \tag{2.20}$$

for $k = 1, \dots, m$, where the M_j's are positive constants satisfying $M_{k+1} \le C_0\,\theta_0^k$ for $k \ge 0$. Note that $\sum_{j=1}^{k+1} M_j \le M_1 + C_0 \sum_{j=1}^{\infty} \theta_0^j \le C_1$, where C_1 is a positive constant independent of ϵ and k. Therefore, by (2.19), we may obtain (2.14), and we complete the proof of Proposition 2.2.

2.1. Proof of Lemma 2.4. By the Fubini theorem, there exists a constant $\alpha_1 \in (0,1)$ such that

$$\epsilon^{\alpha_1} \int_{\partial B_1} \tilde{e}_\epsilon\,(v_\epsilon) \le K_1, \tag{2.21}$$

where $K_1 = K_1(\alpha_1) > 0$ is a universal constant and $B_1 = B_{\epsilon^{\alpha_1}}(a_1)$. Then $\deg(\frac{v_\epsilon}{|v_\epsilon|}, \partial B_1)$ is well defined. We claim that $\deg(\frac{v_\epsilon}{|v_\epsilon|}, \partial B_1) \ne 0$. By contradiction, suppose that $\deg(\frac{v_\epsilon}{|v_\epsilon|}, \partial B_1) = 0$. We introduce the notation $\tilde{E}_\epsilon(u; B)$ as follows:

$$\tilde{E}_\epsilon(u; B) = \int_B \tilde{e}_\epsilon\,(u) \tag{2.22}$$

for B a bounded smooth domain in Ω and for $u \in H^1(B; \mathbb{C})$. Let $\tilde{v}_\epsilon(x) = v_\epsilon(\epsilon^{\alpha_1}\,x + a_1)$ for $x \in \tilde{B}_1$, where $\tilde{B}_1$ is the unit disk in $\mathbb{R}^2$ with its center at the origin. Then $\deg(\frac{\tilde{v}_\epsilon}{|\tilde{v}_\epsilon|}, \partial\tilde{B}_1) = 0$, and $\tilde{v}_\epsilon$ is a minimizer of $\tilde{E}_{\tilde{\epsilon}}(v; \tilde{B}_1)$ for $v \in H^1(\tilde{B}_1; \mathbb{C})$ and $v = \tilde{v}_\epsilon$ on $\partial\tilde{B}_1$, where $\tilde{\epsilon} = \epsilon^{1-\alpha_1}$. Hence, by Theorem 1 of [14],

$$\tilde{E}_{\tilde{\epsilon}}(\tilde{v}_\epsilon; \tilde{B}_1) \le M_3, \tag{2.23}$$

where M_3 is a positive constant independent of ϵ. Moreover, by (2.23), we obtain

$$\tilde{v}_\epsilon \to n_h \quad \text{weakly in} \quad H^1(\tilde{B}_1; \mathbb{C}), \tag{2.24}$$

where $n_h \in H^1(\tilde{B}_1; S^1)$ with zero degree and finite energy.

Now we want to prove that

$$|\nabla\tilde{v}_\epsilon(x)| \le \frac{M_4}{\tilde{\epsilon}} \quad \text{for } |x| \le \tilde{\epsilon}, \tag{2.25}$$

where M_4 is a positive constant independent of ϵ. From (2.23), we have

$$\int_{B_{\tilde{\epsilon}}} \tilde{e}_{\tilde{\epsilon}}\,(\tilde{v}_\epsilon) \le M_3, \tag{2.26}$$

where $B_{\tilde\epsilon}$ is a disk in $\mathbb{R}^2$ with radius $\tilde\epsilon$ and center at the origin. Let $\hat v_\epsilon(x) = \tilde v_\epsilon(\tilde\epsilon\, x)$ for $x \in \tilde B_1$. Then (2.26) implies that

$$\int_{\tilde B_1} \tilde e_1\,(\hat v_\epsilon) \,\le\, M_3. \tag{2.27}$$

Moreover, $\hat v_\epsilon$ is a minimizer of $\tilde E_1(w;\tilde B_1)$ for $w \in H^1(\tilde B_1;\mathbb{C})$ and $w = \hat v_\epsilon$ on $\partial\tilde B_1$. Hence, by [13], we obtain

$$|\nabla \hat v_\epsilon(x)| \,\le\, M_4 \quad \text{ for } x \in \tilde B_1, \tag{2.28}$$

where M_4 is a positive constant independent of ϵ. Thus, by (2.28), we have (2.25).

Since $|v_\epsilon(a_1)| \le \frac12$,

$$|\tilde v_\epsilon(0)| \le \frac12\,. \tag{2.29}$$

Hence, by (2.25) and (2.29), we have

$$\int_{\tilde B_1} \frac{1}{\tilde\epsilon^2}\,(1-|\tilde v_\epsilon|^2)^2 \,\ge c_0, \tag{2.30}$$

where c_0 is a positive constant independent of ϵ. By (2.24), (2.30), and Fatou's lemma, we obtain

$$\liminf_{\tilde\epsilon\to 0+} \tilde E_{\tilde\epsilon}(\tilde v_\epsilon;\tilde B_1) \,\ge\, E(n_h) + c_0, \tag{2.31}$$

where $E(n_h) = \int_{\tilde B_1} \frac12\,|\nabla n_h|^2$ and n_h is of unit length. On the other hand, since $\tilde v_\epsilon$ is a minimizer of $\tilde E_{\tilde\epsilon}(v;\tilde B_1)$ among all $v \in H^1(\tilde B_1;\mathbb{C})$ and $v = \tilde v_\epsilon$ on $\partial\tilde B_1$, a simple comparison yields

$$\begin{aligned} \tilde E_{\tilde\epsilon}(\tilde v_\epsilon;\tilde B_1) &\le \tilde E_{\tilde\epsilon}(n_h;\tilde B_1) + o_\epsilon(1) \\ &= E(n_h) + o_\epsilon(1), \end{aligned} \tag{2.32}$$

as $\epsilon \to 0+$, where $o_\epsilon(1)$ is a small quantity which tends to zero as $\epsilon \to 0+$. Therefore, by (2.31) and (2.32), we obtain a contradiction and complete the proof of Lemma 2.4.

3. Minimization of (1.11). In this section, we prove the following result.

THEOREM 3.1. *Assume $(\psi_\epsilon, u_\epsilon)$ is a minimizer of (1.11) for $\psi \in H^1_\eta(\Omega;\mathbb{C})$, $u \in H^1_g(\Omega;\mathbb{C})$, and $|\psi| = |u|$ in Ω. Then there exist an integer $0 \le N_0 \le \min(d_1, d_2)$, N_0 distinct points $a_1, \dots, a_{N_0}$, $d_1 - N_0$ distinct points $b_1, \dots, b_{d_1-N_0}$, and $d_2 - N_0$ distinct points $c_1, \dots, c_{d_2-N_0}$ in Ω such that*

$$\psi_\epsilon \to \Psi_{a,b} \quad \textit{weakly in } H^1_{loc}(\bar\Omega\backslash\{a_1, \dots, a_{N_0}, b_1, \dots, b_{d_1-N_0}\})$$

and

$$u_\epsilon \to U_{a,c} \quad \textit{weakly in } H^1_{loc}(\bar\Omega\backslash\{a_1, \dots, a_{N_0}, c_1, \dots, c_{d_2-N_0}\}),$$

where

$$\Psi_{a,b}(x) = \prod_{j=1}^{N_0} \frac{x-a_j}{|x-a_j|} \prod_{k=1}^{d_1-N_0} \frac{x-b_k}{|x-b_k|}\, e^{i\,h_{a,b}(x)},$$

$$U_{a,c}(x) = \prod_{j=1}^{N_0} \frac{x-a_j}{|x-a_j|} \prod_{k=1}^{d_2-N_0} \frac{x-c_k}{|x-c_k|}\, e^{i\,h_{a,c}(x)},$$

and $h_{a,b}$ and $h_{a,c}$ are harmonic functions on Ω such that the value of $h_{a,b}$ and $h_{a,c}$ on $\partial\Omega$ is uniquely determined (mod·2π) *by the requirements $\Psi_{a,b} = \eta$ and $U_{a,c} = g$ on $\partial\Omega$, respectively. Note that a_j's, b_k's, and c_l's are all distinct. Moreover,*

(3.1)
$$\begin{aligned} E_\epsilon(\psi_\epsilon, u_\epsilon) \quad &= \pi\,(d_1+d_2)\,\log\tfrac{1}{\epsilon} + W_\eta(a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0}) \\ &\quad + W_g(a_1,\dots,a_{N_0},c_1,\dots,c_{d_2-N_0}) + (d_1+d_2)\,\gamma - 2N_0\,\tilde\gamma + o_\epsilon(1), \end{aligned}$$

where W_η and W_g are renormalized energies (cf. [1]*), and γ and $\tilde\gamma$ are two universal constants.*

Proof of Theorem 3.1. To study the energy minimization and the dynamics of vortices, we consider the cone $\mathbb{C}_0 \equiv \{(s,\psi,u) \in \mathbb{R}^{4,1} : |s| = |\psi| = |u|\}$ in the Minkowski-space $\mathbb{R}^{4,1}$. Note that $\mathbb{C}$ is identified with $\mathbb{R}^2$. Here $\mathbb{R}^{4,1} \simeq \mathbb{R}^5$ is endowed with the Minkowski metric $dx_1^2 + \cdots + dx_4^2 - ds^2$ for $(s,x) \in \mathbb{R}^{4,1}$. We shall denote the upper half cone $\{(s,x) \in \mathbb{C}_0 : s \ge 0\}$ by $\mathbb{C}_0^+$. The metric h on $\mathbb{C}_0$ which is induced from the metric of $\mathbb{R}^{4,1}$ is Riemannian. In fact, one may view $\mathbb{C}_0^+$ as a graph over $\mathbb{R}^4$. Then h in the coordinate system $x \in \mathbb{R}^4$ can be expressed as

$$\begin{aligned} h \quad &= h_{i\,j}(x)\,dx^i \otimes dx^j, \\ h_{i\,j}(x) \quad &= \delta_{ij} - \frac{1}{2}\frac{x_i\,x_j}{|x|^2} \qquad \text{for } i,j \in \{1,2\} \text{ or } \{3,4\}, \\ h_{i\,j}(x) \quad &= 0 \qquad\qquad\qquad\quad \text{otherwise.} \end{aligned} \tag{3.2}$$

We introduce the notation $\|\nabla\,(s,\psi,u)\|^2 = |\nabla\,\psi|^2 + |\nabla\,u|^2 - |\nabla\,s|^2$ and $\|(s,\psi,u)\|^2 = |\psi|^2 + |u|^2 - |s|^2$. Then (1.11) is equivalent to

$$\int_\Omega \frac{1}{2}\,\|\nabla\,(s,\psi,u)\|^2 + \frac{1}{4\,\epsilon^2}\,(1-s^2)^2. \tag{3.3}$$

Let $H^1_*(\Omega;\mathbb{C}_0)$ be the set of all maps $(s,\psi,u) : \Omega \to \mathbb{C}_0$ such that $\int_\Omega \|\nabla\,(s,\psi,u)\|^2 < \infty$ and $\psi = \eta, u = g$ on $\partial\Omega$, where $\eta, g : \partial\Omega \to S^1$ are smooth maps with degrees d_1, d_2, respectively. It is then easy to check that $(s,\psi,u) \in H^1_*(\Omega;\mathbb{C}_0)$ if and only if $(s,\psi,u) : \Omega \to \mathbb{C}_0$, $\psi \in H^1_\eta(\Omega;\mathbb{C}), u \in H^1_g(\Omega;\mathbb{C}), |s| = |\psi| = |u|$ in Ω, and $\int_\Omega |\nabla s|^2 + |\nabla\psi|^2 + |\nabla u|^2 < \infty$. That is, $H^1_*(\Omega;\mathbb{C}_0)$ coincides with the usual H^1-maps from Ω to $\mathbb{C}_0$ when we view $\mathbb{C}_0$ as embedded in $\mathbb{R}^5$. Since the energy functional $\int_\Omega \|\nabla\,(s,\psi,u)\|^2$ is a lower semicontinuous functional, by the direct method of calculus of variations, (3.3) has a minimizer over $H^1_*(\Omega;\mathbb{C}_0)$. Therefore, there is a minimizer $(\psi_\epsilon, u_\epsilon)$ of (1.11) such that $\psi_\epsilon \in H^1_\eta(\Omega;\mathbb{C}), u_\epsilon \in H^1_g(\Omega;\mathbb{C})$, and $|\psi_\epsilon| = |u_\epsilon|$ in Ω.

Now we want to show the energy estimate (3.1). To describe more precisely our results, we let Ω be a bounded smooth domain in $\mathbb{R}^2$ and let $\eta : \partial\Omega \to S^1, g : \partial\Omega \to S^1$ be smooth maps of degrees d_1, d_2, respectively. Note that d_1, d_2 are positive integers. From [1] and [14], it is easy to obtain a pair of maps $w_\epsilon \in H^1_\eta(\Omega,\mathbb{C})$ and $\hat w_\epsilon \in H^1_g(\Omega,\mathbb{C})$ such that $|w_\epsilon| = |\hat w_\epsilon|$ in Ω and

$$E_\epsilon(w_\epsilon, \hat w_\epsilon) \le \pi\,(d_1+d_2)\,\log\frac{1}{\epsilon} + K,$$

where K is a positive constant independent of ϵ. Then we have

$$E_\epsilon(\psi_\epsilon, u_\epsilon) \le E_\epsilon(w_\epsilon, \hat w_\epsilon) \le \pi\,(d_1+d_2)\,\log\frac{1}{\epsilon} + K. \tag{3.4}$$

Hence, by (3.4), Corollary 2.3, and Proposition 2.1, there exist a nonnegative integer $0 \le N_0 \le \min(d_1,d_2)$, N_0 distinct points $a_j, j = 1,\dots,N_0$, $d_1 - N_0$ distinct points

$b_k, k = 1, \dots, d_1 - N_0$, and $d_2 - N_0$ distinct points $c_l, l = 1, \dots, d_2 - N_0$, such that $b_k \neq c_l \ \forall k, l$,

$$\psi_\epsilon \to \Psi_{a,b} \quad \text{weakly in } H^1_{loc}(\bar{\Omega}\backslash\{a_1, \dots, a_{N_0}, b_1, \dots, b_{d_1-N_0}\}), \tag{3.5}$$

and

$$u_\epsilon \to U_{a,c} \quad \text{weakly in } H^1_{loc}(\bar{\Omega}\backslash\{a_1, \dots, a_{N_0}, c_1, \dots, c_{d_2-N_0}\}), \tag{3.6}$$

where $\Psi_{a,b}$ and $U_{a,c}$ are defined by

$$\Psi_{a,b}(x) = \prod_{j=1}^{N_0} \frac{x - a_j}{|x - a_j|} \prod_{k=1}^{d_1-N_0} \frac{x - b_k}{|x - b_k|} e^{i\, h_{a,b}(x)} \tag{3.7}$$

and

$$U_{a,c}(x) = \prod_{j=1}^{N_0} \frac{x - a_j}{|x - a_j|} \prod_{k=1}^{d_2-N_0} \frac{x - c_k}{|x - c_k|} e^{i\, h_{a,c}(x)}. \tag{3.8}$$

Here $h_{a,b}$ and $h_{a,c}$ are H^1-functions on Ω such that the value of $h_{a,b}$ and $h_{a,c}$ on $\partial\Omega$ is uniquely determined (mod$\cdot 2\pi$) by the requirements $\Psi_{a,b} = \eta$ and $U_{a,c} = g$ on $\partial\Omega$, respectively. Note that a_j's, b_k's, and c_l's are all distinct. Moreover, ψ_ϵ has d_1 essential zeros, and u_ϵ has d_2 essential zeros. By the Euler–Lagrange equation of $(\psi_\epsilon, u_\epsilon)$ and the same argument of Proposition 3.3 in [26], we obtain that $h_{a,b}$ and $h_{a,c}$ are harmonic functions on Ω.

A simple computation shows that

$$\begin{aligned} &\tfrac{1}{2} \int_{\Omega\backslash[\cup_{j=1}^{N_0} B_\rho(a_j) \cup_{k=1}^{d_1-N_0} B_\rho(b_k)]} |\nabla\, \Psi_{a,b}|^2(x)\, dx \\ &= \pi\, d_1\, \log \frac{1}{\rho} + W_\eta(a_1, \dots, a_{N_0}, b_1, \dots, b_{d_1-N_0}) + O(\rho) \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} &\tfrac{1}{2} \int_{\Omega\backslash[\cup_{j=1}^{N_0} B_\rho(a_j) \cup_{k=1}^{d_2-N_0} B_\rho(c_k)]} |\nabla\, U_{a,c}|^2(x)\, dx \\ &= \pi\, d_2\, \log \frac{1}{\rho} + W_g(a_1, \dots, a_{N_0}, c_1, \dots, c_{d_2-N_0}) + O(\rho)\,, \end{aligned} \tag{3.10}$$

as $\rho \to 0+$, where W_η and W_g are called the renormalized energies associated with the boundary conditions η and g, respectively (cf. [1]). Hereafter, we set $0 < \epsilon \ll \rho \ll 1$.

Next we want to give a more precise upper bound of $E_\epsilon(\psi_\epsilon, u_\epsilon)$, where E_ϵ is defined in (1.11) and $(\psi_\epsilon, u_\epsilon)$ is the associated minimizer. We will construct a pair of comparison maps $\psi^\epsilon_{a,b} \in H^1_\eta(\Omega; \mathbb{C})$, $u^\epsilon_{a,c} \in H^1_g(\Omega; \mathbb{C})$, and $|\psi^\epsilon_{a,b}| = |u^\epsilon_{a,c}|$ in Ω such that

$$\begin{aligned} E_\epsilon(\psi^\epsilon_{a,b}, u^\epsilon_{a,c}) &= \pi\,(d_1 + d_2)\, \log \frac{1}{\epsilon} + W_\eta(a_1, \dots, a_{N_0}, b_1, \dots, b_{d_1-N_0}) \\ &\quad + W_g(a_1, \dots, a_{N_0}, c_1, \dots, c_{d_2-N_0}) + (d_1 + d_2)\,\gamma \\ &\quad - 2N_0\,\tilde{\gamma} + O(\rho) + o_\epsilon(1), \end{aligned} \tag{3.11}$$

where $\gamma, \tilde{\gamma} > 0$ are universal constants, $O(\rho) \to 0$ as $\rho \to 0$, and $o_\epsilon(1)$ is a small quantity which tends to zero as $\epsilon \to 0$. We may choose ρ sufficiently small such that

$B_\rho(a_j)$'s, $B_\rho(b_k)$'s, and $B_\rho(c_l)$'s are disjoint disks in $\mathbb{R}^2$. The map $\psi^\epsilon_{a,b}$ can be chosen so that

$$\psi^\epsilon_{a,b}(x) = \begin{cases} \Psi_{a,b}(x) & \text{for } |x-a_j|, |x-b_k|, |x-c_l| \geq \rho\,, \\ \psi_0(x-a_j) & \text{for } |x-a_j| \leq \rho/2\,, \\ v_0(x-b_k) & \text{for } |x-b_k| \leq \rho/2\,, \\ |v_0(x-c_l)|\,\Psi_{a,b}(x) & \text{for } |x-c_l| \leq \rho/2\,, \\ \omega_{a,b}(x) & \text{elsewhere,} \end{cases} \tag{3.12}$$

and $u^\epsilon_{a,c}$ can be chosen by

$$u^\epsilon_{a,c}(x) = \begin{cases} U_{a,c}(x) & \text{for } |x-a_j|, |x-b_k|, |x-c_l| \geq \rho\,, \\ u_0(x-a_j) & \text{for } |x-a_j| \leq \rho/2\,, \\ |v_0(x-b_k)|\,U_{a,c}(x) & \text{for } |x-b_k| \leq \rho/2\,, \\ v_0(x-c_l) & \text{for } |x-c_l| \leq \rho/2\,, \\ \omega_{a,c}(x) & \text{elsewhere,} \end{cases} \tag{3.13}$$

where v_0 is the minimizer of $\int_{B_{\rho/2}(0)} e_\epsilon(u)$ with the boundary condition $\frac{x}{|x|}$ on $\partial B_{\rho/2}(0)$, $\omega_{a,b}$ and $\omega_{a,c}$ are the canonical harmonic maps with admissible boundary conditions (cf. [1]), and (ψ_0, u_0) is the minimizer of $\int_{B_{\rho/2}(0)} e_\epsilon(\psi,u)$ for $|\psi| = |u|$ in $B_{\rho/2}(0)$ and $\psi = u = \frac{x}{|x|}$ on $\partial B_{\rho/2}(0)$. Hereafter, $e_\epsilon(u) \equiv \frac{1}{2}\,|\nabla\, u|^2 + \frac{1}{4\,\epsilon^2}\,(1-|u|^2)^2$ and $e_\epsilon(\psi,u) \equiv \frac{1}{2}\,(|\nabla\,\psi|^2 + |\nabla\, u|^2 - |\nabla\,|\psi||^2) + \frac{1}{4\,\epsilon^2}\,(1-|\psi|^2)^2$. By Lemma IX.1 in [1],

$$\int_{B_{\rho/2}(0)} e_\epsilon(v_0) = \pi\,\log\,\frac{\rho}{2\,\epsilon} + \gamma + o_\epsilon(1), \tag{3.14}$$

where γ is a positive universal constant defined by

$$\gamma = \lim_{t\to 0+} I(t) + \pi \log t. \tag{3.15}$$

Here $I(t)$ is defined by

$$I(\epsilon, R) = \min_{v\in V} \int_{B_R(0)} e_\epsilon(v), \text{ and } \quad I(t) = I(t,1) \quad \text{for } \epsilon, R, t > 0, \tag{3.16}$$

where

$$V = \left\{ v : v \in H^1(B_R(0))\,, v = \frac{x}{|x|} \text{ on } \partial B_R(0) \right\}.$$

Now we want to prove

$$\int_{B_{\rho/2}(0)} e_\epsilon(\psi_0, u_0) = 2\,\pi\,\log\,\frac{\rho}{2\,\epsilon} + 2(\gamma - \tilde{\gamma}) + o_\epsilon(1), \tag{3.17}$$

where γ and $\tilde{\gamma}$ are positive universal constants. By the energy comparison, Corollary 2.3, and Proposition 2.1, it is easy to obtain

$$\int_{B_{\rho/2}(0)} e_\epsilon(\psi_0, u_0) = 2\,\pi\,\log\,\frac{\rho}{2\,\epsilon} + O(1). \tag{3.18}$$

To get the delicate estimate of $O(1)$ in (3.18), we define

$$J(\epsilon, R) = \min_{(\psi\,,u)\in W} \int_{B_R(0)} e_\epsilon(\psi, u) \quad \text{and} \quad J(t) = J(t,1) \quad \text{for } \epsilon, R, t > 0\,, \tag{3.19}$$

where

$$W = \left\{ (\psi, u) : \psi\,, u \in H^1(B_R(0))\,, |\psi| = |u| \text{ in } B_R(0)\,, \psi = u = \frac{x}{|x|} \text{ on } \partial B_R(0) \right\}.$$

By scaling, it is obvious that

$$J(\epsilon, R) = J(\epsilon/R) = J(1, R/\epsilon)\,. \tag{3.20}$$

As for the proof of Lemma III.1 in [1], we have

$$J(t_1) \le J(t_2) + 2\pi \log(t_2/t_1) \quad \forall t_1 \le t_2; \tag{3.21}$$

i.e., the function $J(t) + 2\pi \log t$ is nondecreasing. Let $\gamma_1 = \lim_{t\to 0+} J(t) + 2\pi \log t$. Then it is easy to check that $\gamma_1 < 2\gamma$, and we may set $\gamma_1 = 2(\gamma - \tilde{\gamma})$, where $\tilde{\gamma}$ is a positive universal constant. Here we have used the fact that

$$\begin{aligned} \int_{B_{\rho/2}(0)} e_\epsilon(\psi_0, u_0) &\le \int_{B_{\rho/2}(0)} e_\epsilon(v_0, v_0) \\ &= 2 \int_{B_{\rho/2}(0)} e_\epsilon(v_0) - \int_{B_{\rho/2}(0)} \frac{1}{2} |\nabla\, |v_0||^2 + \frac{1}{4\,\epsilon^2} \left(1 - |v_0|^2\right)^2 dx \\ &= 2\,\pi\, \log \frac{\rho}{2\,\epsilon} + 2(\gamma - \gamma_2) + o_\epsilon(1), \end{aligned}$$

where

$$\gamma_2 = \lim_{\epsilon \longrightarrow 0} \frac{1}{2} \int_{B_{\rho/2}(0)} \frac{1}{2} |\nabla\, |v_0||^2 + \frac{1}{4\,\epsilon^2} \left(1 - |v_0|^2\right)^2 dx.$$

By Theorem 11.1 of [20], $v_0 = f_\epsilon(r)\, e^{i\,\theta}$ is the radial solution of the Ginzburg–Landau equation, where f_ϵ satisfies an ordinary differential equation. The quantitative properties of f_ϵ can be found in [3], [9], and [10]. Hence γ_2 is a universal positive constant, and we may complete the proof of (3.17).

By (3.12) and (3.13), we have

$$\int_{B_{\rho/2}(a_j)} e_\epsilon(\psi^\epsilon_{a,b}, u^\epsilon_{a,c}) = \int_{B_{\rho/2}(0)} e_\epsilon(\psi_0, u_0), \tag{3.22}$$

$$\int_{B_{\rho/2}(b_k)} e_\epsilon(\psi^\epsilon_{a,b}, u^\epsilon_{a,c}) = \int_{B_{\rho/2}(0)} e_\epsilon(v_0) + \int_{B_{\rho/2}(0)} |v_0|^2\, |\nabla\, U_{a,c}(x + b_k)|^2\,, \tag{3.23}$$

and

$$\int_{B_{\rho/2}(c_l)} e_\epsilon(\psi^\epsilon_{a,b}, u^\epsilon_{a,c}) = \int_{B_{\rho/2}(0)} e_\epsilon(v_0) + \int_{B_{\rho/2}(0)} |v_0|^2\, |\nabla\, \Psi_{a,b}(x + c_l)|^2\,. \tag{3.24}$$

Hence by (3.9), (3.10), (3.14), (3.17), (3.22)–(3.24), and the results of [1], we may obtain (3.11) and

$$\begin{aligned} E_\epsilon(\psi_\epsilon, u_\epsilon) \le{}& \pi\,(d_1+d_2)\,\log\frac{1}{\epsilon} + W_\eta(a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0}) \\ &+ W_g(a_1,\dots,a_{N_0},c_1,\dots,c_{d_2-N_0}) + (d_1+d_2)\,\gamma - 2N_0\,\tilde\gamma + O(\rho) \\ &+ o_\epsilon(1). \end{aligned} \tag{3.25}$$

By (3.5), (3.6), (3.9), (3.10), (3.14), (3.17), and Fatou's lemma, we can derive that

$$\begin{aligned} E_\epsilon(\psi_\epsilon, u_\epsilon) \ge{}& \pi\,(d_1+d_2)\,\log\frac{1}{\epsilon} + W_\eta(a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0}) \\ &+ W_g(a_1,\dots,a_{N_0},c_1,\dots,c_{d_2-N_0}) + (d_1+d_2)\,\gamma - 2N_0\,\tilde\gamma + O(\rho) \\ &+ o_\epsilon(1). \end{aligned} \tag{3.26}$$

Therefore, by (3.25), (3.26), and suitable choice of ρ, we obtain (3.1), and we complete the proof of Theorem 3.1.

Remark 3.1. From (3.1), N_0, a_k's, b_j's, and c_l's are determined by the minimization of

$$W_\eta(a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0}) + W_g(a_1,\dots,a_{N_0},c_1,\dots,c_{d_2-N_0}) + (d_1+d_2)\,\gamma - 2N_0\,\tilde\gamma$$

for $N_0 \ge 0$, and $a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0},c_1,\dots,c_{d_2-N_0}$ are distinct points in Ω. Hence they may depend on η, g, and Ω. Now we divide our discussion of N_0 into two cases as follows.

Case (a). Suppose $\eta = g$ and $d_1 = d_2$. Then it is obvious that $N_0 = d_1 = d_2$ and $(a_1,\dots,a_{N_0})$ is the global minimal point of W_η.

Case (b). Suppose η is quite different from g. Then N_0 may become zero. We may give an example for the case that $N_0 = 0 < \min\{d_1, d_2\}$. Assume $d_1 = d_2 = 1$ and Ω is a smooth dumbbell-shaped domain defined by $\Omega = B_1((-\xi,0)) \cup B_1((\xi,0)) \cup D_{r_0}$, where $\xi \ge 3$ is a constant, $D_{r_0} \subset [-\xi+\frac{1}{2}, \xi+\frac{1}{2}] \times [-r_0, r_0]$ is the neck region joining $B_1((-\xi,0))$ and $B_1((\xi,0))$, and r_0 is a positive constant. Let

$$g(x) = \begin{cases} e^{i\,\arg(x-(-\xi,0))} & \text{for } x \in \partial B_1((-\xi,0))\backslash\partial D_{r_0}\,, \\ e^{i\,\tilde g(x)} & \text{for } x \in \partial[B_1((\xi,0)) \cup D_{r_0}]\backslash\partial B_1((-\xi,0))\,, \end{cases}$$

and

$$\eta(x) = \begin{cases} e^{i\,\arg(x-(\xi,0))} & \text{for } x \in \partial B_1((\xi,0))\backslash\partial D_{r_0}\,, \\ e^{i\,\tilde\eta(x)} & \text{for } x \in \partial[B_1((-\xi,0)) \cup D_{r_0}]\backslash\partial B_1((\xi,0))\,, \end{cases}$$

where both $\tilde g$ and $\tilde\eta$ are smooth and monotone functions such that $\deg(g, \partial\Omega) = \deg(\eta, \partial\Omega) = 1$, respectively. From the definition of the renormalized energy (see [1]), it is easy to check that, as $0 < r_0 \ll 1$,

$$W_g(a) + W_\eta(b) \ll \min_{c\in\Omega} W_g(c) + W_\eta(c),$$

for a is close to $(-\xi, 0)$ and b is close to $(\xi, 0)$. Hence N_0 must be zero.

4. Vortex dynamics in *p*-wave superconductivity. In this section, we consider the dynamics of vortices in the gradient flow of (1.11). There are basically two methods for deriving these dynamical laws. The first one that applies to our problem is somewhat restrictive in the range of parameters for regularization, though it is less restrictive on the initial data. This first method was used by the first author in the heat flow of Ginzburg–Landau vortices (see Lecture 3 in [14] and the references therein). The second method was used by Colliander–Jerrard [4] and by Lin and Xin [18] in the study of dynamics of Ginzburg–Landau–Schrödinger vortices; see also [16]. We shall first sketch a proof of Theorem 4.1 by using the first method and then give a more detailed proof by using the second method.

The gradient flow of (1.11) is given by

(4.1)
$$\begin{cases} \partial_t w = -\text{ grad } \tilde{E}_\epsilon(w), & w(x,t) = (s,\psi,u)(x,t) \in \mathbb{C}_0^+ & \text{for } x \in \Omega, t > 0\,, \\ s = 1\,, \psi = \eta\,, & u = g & \text{for } x \in \partial\Omega\,, t > 0\,, \\ s = |\psi_0|\,, \psi = \psi_0\,, & u = u_0 & \text{for } x \in \Omega\,, t = 0\,, \end{cases}$$

where $\tilde{E}_\epsilon(w) \equiv \int_\Omega \frac{1}{2}\|\nabla w\|^2 + \frac{1}{4\,\epsilon^2}\,(1-s^2)^2$. Note that $\tilde{E}_\epsilon$ is equivalent to (1.11) for w is a $\mathbb{C}_0^+$-valued map. Hereafter, ψ_0 and u_0 are smooth maps from Ω to $\mathbb{C}$, ψ_0 has d_1 essential zeros (degree ± 1 vortices) in Ω, and u_0 has d_2 essential zeros (degree ± 1 vortices) in Ω such that $|\psi_0| = |u_0|$ in Ω, and

$$E_\epsilon(\psi_0, u_0) \le \pi\,(d_1 + d_2)\,\log\frac{1}{\epsilon} + \chi_1, \quad \int_\Omega |\nabla |u_0||^2\,dx \le \chi_1, \tag{4.2}$$

where χ_1 is a positive constant independent of ϵ. It is not obvious that (4.1) has a unique weak solution. The main difficulties are that the target manifold $\mathbb{C}_0^+$ may have positive intrinsic curvature somewhere and the metric h on $\mathbb{C}_0^+$ is singular at the vertex $(0,0)$. To overcome this, we approximate $\mathbb{C}_0^+$ by a family of smooth graphs $\{\mathbb{C}_\delta\}$ over $\mathbb{R}^4$ in $\mathbb{R}^{4,1}$. Due to the smoothness of $\mathbb{C}_\delta$, we obtain the regular solution of the gradient flow (4.1) on the target manifold $\mathbb{C}_\delta$. Such a regular solution can be regarded as an approximated solution of the gradient flow (4.1) on the target manifold $\mathbb{C}_0^+$. Hereafter, $\mathbb{C}_\delta$ is one of the sheets of $s^2 = |\psi|^2 + \frac{1}{2}\,\delta^2 = |u|^2 + \frac{1}{2}\,\delta^2$ which lies in $\{s > 0\}$, where $\delta > 0$ is a small parameter. The induced metric on $\mathbb{C}_\delta$ in $\mathbb{R}^{4,1}$ is given by

$$\begin{aligned} h_\delta &= \text{diag } (h_\delta^1, h_\delta^2), \\ h_\delta^1 &= \left(1 - \frac{1}{2}\,\frac{r^2}{r^2+\delta^2}\right) dr^2 + r^2\,d\theta_1^2, \\ h_\delta^2 &= \left(1 - \frac{1}{2}\,\frac{r^2}{r^2+\delta^2}\right) dr^2 + r^2\,d\theta_2^2, \end{aligned} \tag{4.3}$$

where (r,θ_1) is the polar coordinate of $\psi \in \mathbb{R}^2$ and (r,θ_2) is the polar coordinate of $u \in \mathbb{R}^2$.

In our first approach, we shall set $\delta = \epsilon$ and use $\mathbb{C}_\epsilon$ to approximate $\mathbb{C}_0$. Then we consider the gradient flow (4.1) on the target manifold $\mathbb{C}_\epsilon$ as the approximated gradient flow given by

(4.4)
$$\begin{cases} \partial_t w = -\text{ grad } \tilde{E}_\epsilon(w), & w(x,t) = (s,\psi,u)(x,t) \in \mathbb{C}_\epsilon & \text{for } x \in \Omega, t > 0, \\ s = 1 + \frac{1}{2}\epsilon^2, \psi = \eta\,, & u = g & \text{for } x \in \partial\Omega, t > 0, \\ s = s_0, \psi = \psi_0, & u = u_0 & \text{for } x \in \Omega, t = 0, \end{cases}$$

where $s_0 = (|\psi_0|^2 + \frac{1}{2}\epsilon^2)^{\frac{1}{2}}$. Note that $(s_0, \psi_0, u_0)(x) \in \mathbb{C}_\epsilon$ for $x \in \Omega$. To study the dynamics of vortices in (4.4), we may rescale the time variable by $\lambda_\epsilon = \log(1/\epsilon)$. Then (4.4) becomes

(4.5)
$$\begin{cases} \lambda_\epsilon^{-1}\,\partial_t\, w = -\text{ grad }\tilde{E}_\epsilon(w), & w(x,t) = (s,\psi,u)(x,t) \in \mathbb{C}_\epsilon & \text{for } x \in \Omega, t > 0, \\ s = 1 + \frac{1}{2}\epsilon^2\,, \psi = \eta\,, & u = g & \text{for } x \in \partial\Omega\,, t > 0, \\ s = s_0\,, \psi = \psi_0\,, & u = u_0 & \text{for } x \in \Omega\,, t = 0. \end{cases}$$

From [5], (4.5) has a regular solution w_ϵ with $\|\nabla w_\epsilon(\cdot,t)\|_{L^\infty} \le C/\epsilon$ for almost all $t > 0$, where C is a positive constant.

By (4.5), it is easy to check that

$$\frac{d}{d\,t}\,\tilde{E}_\epsilon(w_\epsilon) = -\left(\log\frac{1}{\epsilon}\right)^{-1}\int_\Omega \|\partial_t\, w_\epsilon\|^2\,dx \tag{4.6}$$

for $t > 0$, where $\|\cdot\|$ is the norm of $\mathbb{R}^{4,1}$ defined in section 3. Then by (4.2), (4.6), and $|\psi_0| = |u_0|$ in Ω, we have

$$\tilde{E}_\epsilon(w_\epsilon)(t) \le \tilde{E}_\epsilon(s_0,\psi_0,u_0) \le \pi\,(d_1+d_2)\,\log\frac{1}{\epsilon} + 2\chi_1 \tag{4.7}$$

for $t > 0$. Hence we have

$$E_\epsilon(\psi_\epsilon, u_\epsilon)(t) \le (d_1+d_2)\,\log\frac{1}{\epsilon} + 3\chi_1 \tag{4.8}$$

for $t > 0$. Note that we have used the fact that

$$\|\nabla\, w_\epsilon\|^2 \ge \|\nabla\,(|u_\epsilon|,\psi_\epsilon,u_\epsilon)\|^2\,,\quad \frac{1}{\epsilon^2}\,(1-s_\epsilon^2)^2 \ge \frac{1}{\epsilon^2}\,(1-|u_\epsilon|^2)^2 - 1, \tag{4.9}$$

where $w_\epsilon = (s_\epsilon, \psi_\epsilon, u_\epsilon), |\psi_\epsilon| = |u_\epsilon|$, and $s_\epsilon = (|u_\epsilon|^2 + \frac{1}{2}\epsilon^2)^{\frac{1}{2}}$. Thus, by (4.8) and Corollary 2.3,

$$E_\epsilon(\psi_\epsilon)(t) \le \pi\,d_1\,\log\frac{1}{\epsilon} + O(1)\,,\quad E_\epsilon(u_\epsilon)(t) \le \pi\,d_2\,\log\frac{1}{\epsilon} + O(1) \tag{4.10}$$

for $t > 0$, where $O(1)$ is a bounded quantity depending only on χ_1. Therefore, by Proposition 2.1 and arguments in Lecture 3 of [14], there exist an integer N_0, $0 \le N_0 \le \min(d_1, d_2)$, N_0 distinct points $a_j(t)$, $j = 1,\dots,N_0$, $d_1 - N_0$ distinct points $b_k(t)$, $k = 1,\dots,d_1 - N_0$, and $d_2 - N_0$ distinct points $c_l(t)$, $l = 1,\dots,d_2 - N_0$, such that, for a subsequence of $\epsilon \to 0$, the essential zeros of ψ_ϵ tend to a_j's and b_k's; the essential zeros of u_ϵ tend to a_j's and c_l's. Moreover, for this subsequence of $\epsilon \to 0$, one has

$$\begin{aligned} &\psi_\epsilon(\cdot,t) \to \Psi_{a,b}(\cdot,t) \quad \text{weakly in } H^1_{loc}(\bar{\Omega}\backslash\{a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0}\}), \\ &u_\epsilon(\cdot,t) \to U_{a,c}(\cdot,t) \quad \text{weakly in } H^1_{loc}(\bar{\Omega}\backslash\{a_1,\dots,a_{N_0},c_1,\dots,c_{d_2-N_0}\}) \end{aligned} \tag{4.11}$$

for $0 \le t \le T$, where $T > 0$ is chosen so that all a_j's, b_k's, and c_l's stay inside Ω and for $0 \le t \le T$, no collision occurs on a_j's, b_k's, and c_l's. Here

$$\Psi_{a,b}(x,t) \equiv \prod_{j=1}^{N_0}\frac{x-a_j(t)}{|x-a_j(t)|}\prod_{k=1}^{d_1-N_0}\frac{x-b_k(t)}{|x-b_k(t)|}\,e^{i\,h_{a,b}(x,t)}$$

and

$$U_{a,c}(x,t) \equiv \prod_{j=1}^{N_0} \frac{x-a_j(t)}{|x-a_j(t)|} \prod_{l=1}^{d_2-N_0} \frac{x-c_l(t)}{|x-c_l(t)|}\, e^{i\, h_{a,c}(x,t)},$$

where $h_{a,b}$ and $h_{a,c}$ are H^1-functions with bounded H^1-norms independent of ϵ and t.

We remark that the existence of a subsequence of $\epsilon \to 0$ so that (4.11) is valid for all $0 \le t \le T$ (not just for one t that will be an easy consequence of Proposition 2.1) is probably the key point in this step. It follows in a manner identical to that in Lecture 3 of [14]. Note, however, that N_0 may depend on t in this stage. Indeed, we let μ_ϵ be a Radon measure defined by

$$\mu_\epsilon(t) = \left(\log\frac{1}{\epsilon}\right)^{-1} \left[\frac{1}{2}\,\|\nabla w_\epsilon\|^2 + \frac{1}{4\,\epsilon^2}\,(1-s_\epsilon^2)^2\right] dx. \tag{4.12}$$

For $\phi \in C_0^1(\mathbb{R}^2)$, we obtain

(4.13)

$$\begin{aligned}
\frac{d}{d\,t}\int_\Omega \phi^2(x)\,\mu_\epsilon(t) &= -\left(\log\frac{1}{\epsilon}\right)^{-2}\int_\Omega \phi^2\,\|\partial_t w_\epsilon\|^2 - \left(\log\frac{1}{\epsilon}\right)^{-1}\int_\Omega 2\phi\,\nabla\phi\,\nabla w_\epsilon\,\cdot\,\partial_t w_\epsilon \\
&\le \int_\Omega \phi^2\cdot\mu_\epsilon(t) + \left(\log\frac{1}{\epsilon}\right)^{-1} C(\phi)\int_\Omega \|\partial_t w_\epsilon\|^2 \\
&\le C(\phi)\,[\|\mu_\epsilon(0)\| + K_\epsilon'(t)],
\end{aligned}$$

where $C(\phi)$ is a positive constant depending on the C^1-norm of ϕ,

$$K_\epsilon(t) = \left(\log\frac{1}{\epsilon}\right)^{-1}\int_0^t\int_\Omega \|\partial_t w_\epsilon\|^2,$$

and $\|\mu_\epsilon(0)\|$ denotes the total measures of $\mu_\epsilon(0)$. Here we have used (4.5), (4.6), and integration by parts. Hence by Lecture 3 of [14] or the argument of [15] (Proof of Theorem 2.1(iii)), we may obtain such a subsequence so that (4.11) is valid and $a_j(t)$'s, $b_k(t)$'s, and $c_l(t)$'s are continuous in t.

For the dynamics of vortices, we have the equations of a_j's, b_k's, and c_l's as follows.

THEOREM 4.1. *Assume a_j's, b_k's, and c_l's are as above. Then they satisfy a system of ordinary differential equations given by*

(4.14)

$$\begin{cases}
-2\,\dot a_j &= \nabla_{a_j} W_\eta(a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0}) + \nabla_{a_j} W_g(a_1,\dots,a_{N_0},c_1,\dots,c_{d_2-N_0}),\\
-\dot b_k &= \nabla_{b_k} W_\eta(a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0}),\\
-\dot c_l &= \nabla_{c_l} W_g(a_1,\dots,a_{N_0},c_1,\dots,c_{d_2-N_0})
\end{cases}$$

for $j = 1,\dots,N_0, k = 1,\dots,d_1-N_0$, and $l = 1,\dots,d_2-N_0$, where W_η and W_g are the renormalized energies (cf. [1]*). Here $\nabla_{a_j}, \nabla_{b_k}$, and ∇_{c_l} are to take the gradient only on a_j, b_k, and c_l, respectively.*

Hereafter, we regard N_0 as a fixed constant with respect to time t, $0 \le t \le T$. Actually, such an assumption is reasonable. For instance, suppose the energy of the initial data (s_0,ψ_0,u_0) satisfies that $\tilde E_\epsilon(s_0,\psi_0,u_0)$ is equal to the right side of (3.1).

Since the energy is dissipative with time t, then it is impossible to decrease N_0 as time t increases. Moreover, we may claim that N_0 may not increase with time t, generically. From (4.14) and the standard theorem of ordinary differential equations, the map from

$$\Gamma(0) \equiv (a_1(0), \dots, a_{N_0}(0), b_1(0), \dots, b_{d_1-N_0}(0), c_1(0), \dots, c_{d_2-N_0}(0)) \in \mathbb{R}^{2(d_1+d_2)}$$

to

$$\Gamma(t) \equiv (a_1(t), \dots, a_{N_0}(t), b_1(t), \dots, b_{d_1-N_0}(t), c_1(t), \dots, c_{d_2-N_0}(t)) \in \mathbb{R}^{2(d_1+d_2)}$$

is a diffeomorphism for $t > 0$. Hence the collision manifold $\Lambda_{j,k}(t) \equiv \{\Gamma(0) : b_j(t) = c_k(t)\}$ has codimension two for $t > 0$. Thus $\cup_{j,k,t>0} \Lambda_{j,k}(t)$ has measure zero in $\mathbb{R}^{2(d_1+d_2)}$. Therefore, the trajectories of b_j's and c_k's may not come together, and N_0 may not increase with time t, generically.

Proof of Theorem 4.1 (sketch).

By (4.2), (4.6), (4.9), and Theorem 3.1, we have

$$\begin{aligned} \left(\log \frac{1}{\epsilon}\right)^{-1} \int_0^T \int_\Omega \|\partial_t w_\epsilon\|^2 &= -\int_0^T \frac{d}{d\,t}\, \tilde{E}_\epsilon(w_\epsilon)\, dt \\ &= \tilde{E}_\epsilon(s_0, \psi_0, u_0) - \tilde{E}_\epsilon(w_\epsilon) \\ &\le C(\eta, g, \chi_1, \Omega). \end{aligned} \tag{4.15}$$

Since a_j's, b_k's, and c_l's are distinct, we may choose a small constant $\delta_* > 0$ such that $B_R(a_j)$'s, $B_R(b_k)$'s, and $B_R(c_l)$'s are disjoint for $R \in [\delta_*/2, \delta_*]$. We multiply (4.5) by ∇w_ϵ and then integrate on $B_R(\beta)$, where β is one of a_j's, b_k's, and c_l's. Using integration by parts, we have

$$\begin{aligned} &\frac{-1}{\log \frac{1}{\epsilon}} \int_{B_R(\beta)} \langle \partial_t w_\epsilon\,, \nabla w_\epsilon \rangle \\ &= \frac{1}{4\,\epsilon^2} \int_{\partial B_R(\beta)} (1 - s_\epsilon^2)^2\, \nu + \frac{1}{2} \int_{\partial B_R(\beta)} \|\nabla w_\epsilon\|^2\, \nu - \int_{\partial B_R(\beta)} \left\langle \frac{\partial w_\epsilon}{\partial \nu}\,, \nabla w_\epsilon \right\rangle, \end{aligned} \tag{4.16}$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product in the hyperbolic space $\mathbb{R}^{4,1}$. On the other hand, we calculate with $e_\epsilon \equiv \frac{1}{2}\, \|\nabla\, w_\epsilon\|^2 + \frac{1}{4\,\epsilon^2}\, (1 - s_\epsilon^2)^2$ that

$$\begin{aligned} &\frac{1}{\log \frac{1}{\epsilon}}\, \frac{d}{d\,t} \int_{B_R(\beta)} x \cdot e_\epsilon(w_\epsilon)\, dx \\ &= -\frac{1}{(\log \frac{1}{\epsilon})^2} \int_{B_R(\beta)} x\, \|\partial_t w_\epsilon\|^2\, dx - \frac{1}{\log \frac{1}{\epsilon}} \int_{B_R(\beta)} \langle \partial_t w_\epsilon\,, \nabla w_\epsilon \rangle\, dx \\ &\quad + \frac{1}{\log \frac{1}{\epsilon}} \int_{\partial B_R(\beta)} x \left\langle \frac{\partial w_\epsilon}{\partial \nu}\,, \partial_t w_\epsilon \right\rangle. \end{aligned} \tag{4.17}$$

For (4.16) and (4.17), we may consider the isometric embedding from $\mathbb{C}_\epsilon$ to $\mathbb{R}^k$, where k is a positive integer, and we transform (4.5) into a standard nonlinear parabolic system. Then using such a nonlinear parabolic system and integration by parts, we may obtain (4.16) and (4.17), respectively.

In order to derive dynamical equations for vortices from (4.17), we have to establish the strong convergence in H^1 of maps away from vortices. For the strong

convergence theorems in section 5 of [15], it is crucial to get Pohozaev's identity for the steady state equation of (4.5) given by

$$-\text{ grad }\tilde{E}_\epsilon(w)=0 \quad \text{ for } x\in\Omega\,, \tag{4.18}$$

with boundary conditions $s=1+\frac{1}{2}\epsilon^2, \psi=\eta, u=g$ on $\partial\Omega$. By the isometric embedding from $\mathbb{C}_\epsilon$ into $\mathbb{R}^k$, where k is a positive integer, we may transform (4.18) into a semilinear elliptic system. Hence it is easy to obtain the associated Pohozaev identity given by

$$\begin{aligned}&\int_{B_r(0)}\frac{1}{4\,\epsilon^2}\,(1-s_\epsilon^2)^2+\frac{r}{2}\int_{\partial B_r(0)}\left\|\frac{\partial w_\epsilon}{\partial\nu}\right\|^2\\&=\frac{r}{4\,\epsilon^2}\int_{\partial B_r(0)}(1-s_\epsilon^2)^2+\frac{r}{2}\int_{\partial B_r(0)}\left\|\frac{\partial w_\epsilon}{\partial\tau}\right\|^2\end{aligned} \tag{4.19}$$

for $0<r<1$, where $\frac{\partial}{\partial\nu}$ and $\frac{\partial}{\partial\tau}$ denote normal and tangential gradients, respectively. Hence by the same argument of Lemma 5.4 in [15], we may have the associated strong convergence theorems for (4.5). Therefore, by (4.15)–(4.17) and the associated strong convergence theorems for (4.5), we follow the argument of Theorem 2.1(iii) in [15], and we may complete the proof of Theorem 4.1.

Next we shall give an alternate approach to the vortex dynamics. We set $\delta=\delta(\epsilon)$ such that $0<\delta\ll\epsilon\ll 1$, and we use $\mathbb{C}_\delta$ to approximate $\mathbb{C}_0$. This is more accurate than using $\mathbb{C}_\epsilon$ to approximate $\mathbb{C}_0$. Then we consider the gradient flow (4.1) on the target manifold $\mathbb{C}_\delta$ as the approximated gradient flow given by

$$\begin{cases}\partial_t\, w=-\text{ grad }\tilde{E}_\epsilon(w), & w(x,t)=(s,\psi,u)(x,t)\in\mathbb{C}_\delta & \text{for } x\in\Omega, t>0\,,\\ s=1+\frac{1}{2}\delta^2\,,\psi=\eta\,, & u=g & \text{for } x\in\partial\Omega\,, t>0\,,\\ s=s_0\,,\psi=\psi_0\,, & u=u_0 & \text{for } x\in\Omega\,, t=0\,,\end{cases} \tag{4.20}$$

where $s_0=(|\psi_0|^2+\frac{1}{2}\delta^2)^{\frac{1}{2}}$. Hereafter, ψ_0 and u_0 are smooth maps from Ω to $\mathbb{C}$, ψ_0 has d_1 essential zeros (vortices) at a_j^0's and b_k^0's in Ω, and u_0 has d_2 essential zeros (vortices) at a_j^0's and c_l^0's in Ω for $j=1,\dots,N_0, k=1,\dots,d_1-N_0, l=1,\dots,d_2-N_0$, such that $|\psi_0|=|u_0|$ in Ω, and

$$\begin{aligned}\tilde{E}_\epsilon(s_0,\psi_0,u_0)=\pi\,(d_1+d_2)\,\log\frac{1}{\epsilon}+W_\eta(a_1^0,\dots,a_{N_0}^0,b_1^0,\dots,b_{d_1-N_0}^0)\\+W_g(a_1^0,\dots,a_{N_0}^0,c_1^0,\dots,c_{d_2-N_0}^0)+(d_1+d_2)\,\gamma-2N_0\,\tilde{\gamma}+o_\epsilon(1),\end{aligned} \tag{4.21}$$

where γ and $\tilde{\gamma}$ are two universal constants defined in Theorem 3.1. Note that $(s_0,\psi_0,u_0)(x)\in\mathbb{C}_\delta$ for $x\in\Omega$. To study the dynamics of vortices in (4.20), we may rescale the time variable by $\lambda_\epsilon=\log(1/\epsilon)$. Then (4.20) becomes

$$\begin{cases}\lambda_\epsilon^{-1}\,\partial_t\, w=-\text{ grad }\tilde{E}_\epsilon(w), & w(x,t)=(s,\psi,u)(x,t)\in\mathbb{C}_\delta & \text{for } x\in\Omega, t>0\,,\\ s=1+\frac{1}{2}\delta^2\,,\psi=\eta\,, & u=g & \text{for } x\in\partial\Omega\,, t>0\,,\\ s=s_0\,,\psi=\psi_0\,, & u=u_0 & \text{for } x\in\Omega\,, t=0\,.\end{cases} \tag{4.22}$$

From [5], (4.22) has a regular solution $w_{\epsilon,\delta}$ satisfying

$$\|\nabla w_{\epsilon,\delta}(\cdot,t)\|_{L^\infty}\le C/\delta. \tag{4.23}$$

Since $0 < \delta = \delta(\epsilon) \ll \epsilon \ll 1$, the inequality (4.23) cannot ensure $\|\nabla w_\epsilon(\cdot,t)\|_{L^\infty} \le C/\epsilon$ for all $t > 0$, where C is a positive constant. Hereafter, we denote $w_{\epsilon,\delta}$ as w_ϵ for notation convenience. Hence the strong convergence theorem may not be true for (4.22), and we cannot follow the argument of Theorem 4.1 to derive the dynamics of vortices. To overcome such a difficulty, we may follow the idea of energy concentration (cf. [15] and [16]) and obtain (4.24). Then we use the energy comparison and the Gronwall inequality (cf. [16, pp. 746–754]) to prove Theorem 4.2. Such an idea can also be found in [4] and [18].

THEOREM 4.2. *Suppose the initial condition* (s_0, ψ_0, u_0) *satisfies* (4.21) *and* $(s_0, \psi_0, u_0)(x) \in \mathbb{C}_\delta$ *for* $x \in \Omega$. *Moreover,* a_j^0*'s and* b_k^0*'s are distinct essential zeros of* ψ_0*, and* a_j^0*'s and* c_l^0*'s are distinct essential zeros of* u_0*, respectively. Let* $w_\epsilon = (s_\epsilon, \psi_\epsilon, u_\epsilon)$ *be the solution of* (4.22). *Then* (4.11) *holds for* ψ_ϵ *and* u_ϵ*, and* (4.14) *is also valid.*

The target manifold of Theorem 4.1 is $\mathbb{C}_\epsilon$, and the target manifold of Theorem 4.2 is $\mathbb{C}_\delta\,, 0 < \delta \ll \epsilon$. This is the main difference between Theorems 4.1 and 4.2. Moreover, the target manifold $\mathbb{C}_\delta$ of Theorem 4.2 is closer to $\mathbb{C}_0$ than the target manifold $\mathbb{C}_\epsilon$ of Theorem 4.1. However, the dynamics of vortices in Theorem 4.1 is same as the dynamics of vortices in Theorem 4.2. Thus it is reasonable to say that (4.14) is the motion equation of vortices in the gradient flow (4.1) on the target manifold $\mathbb{C}_0$.

Proof of Theorem 4.2. As for the proof of Theorem 4.1, we may obtain (4.11) for the solution w_ϵ of (4.22). In addition, $\psi_\epsilon(\cdot,t)$ has distinct essential zeros near $a_j(t)$'s and $b_k(t)$'s, and $u_\epsilon(\cdot,t)$ has distinct essential zeros near $a_j(t)$'s and $c_l(t)$'s for $j = 1,\dots,N_0\,, k = 1,\dots,d_1 - N_0\,, l = 1,\dots,d_2 - N_0$. Furthermore, (4.15)–(4.17) also hold for the solution w_ϵ of (4.22). We want to derive the system of ordinary differential equations for $a_j(t)$'s, $b_k(t)$'s, and $c_l(t)$'s. For notation convenience, we set

$$\begin{aligned}\vec{a} &= (a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0},c_1,\dots,c_{d_2-N_0})\,,\\ W(\vec{a}) &= W_\eta(a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0}) + W_g(a_1,\dots,a_{N_0},c_1,\dots,c_{d_2-N_0}).\end{aligned}$$

Let $a(t)$ be any one component of $\vec{a}(t)$. For any $t_0 \ge 0$, we let $R > 0$ be a suitable constant so that, for t close to t_0, $a(t) \in B_{R/2}(a(t_0))$ by the continuity of $a(t)$ in t. From [15] and [16], we have

$$\int_{B_R(a(t_0))} \frac{1}{\log\frac{1}{\epsilon}}\, x \cdot e_\epsilon(w_\epsilon)\,dx \to \xi(a)\,a(t) \quad \text{as } \epsilon = \epsilon_n \to 0+\,, \tag{4.24}$$

where

$$\xi(a) = \begin{cases} 2 & \text{if} \quad a \in \{a_1,\dots,a_{N_0}\}\,,\\ 1 & \text{if} \quad a \in \{b_1,\dots,b_{d_1-N_0},c_1,\dots,c_{d_2-N_0}\}\,.\end{cases}$$

As for [16, pp. 747], we may use a family of approximation of the identity to convolute both sides of (4.17). Then, by (4.11), (4.15), (4.16), and (4.24), $\dot{a}(t)$ is bounded in t. Moreover, by (2.28) of [16], we may obtain

$$\begin{aligned}\xi(a)^2\,|\dot{a}(t)|^2 &= \left|\frac{1}{\pi\,\log\frac{1}{\epsilon}}\int_{B_R(a(t_0))} \langle \partial_t w_\epsilon\,, \nabla w_\epsilon\rangle\right|^2 + o_\epsilon(1)\\ &\le \frac{1}{\pi\,\log\frac{1}{\epsilon}}\int_{B_R(a(t_0))} \|\partial_t\, w_\epsilon\|^2 + o_\epsilon(1)\,.\end{aligned} \tag{4.25}$$

Hence, by (4.15) and (4.25), we have

(4.26)
$$\int_0^T \sum_{j=1}^{N_0} 2|\dot{a}_j|^2 + \sum_{k=1}^{d_1-N_0} |\dot{b}_k|^2 + \sum_{l=1}^{d_2-N_0} |\dot{c}_l|^2 \;\le \left(\log\frac{1}{\epsilon}\right)^{-1} \int_0^T \int_\Omega \|\partial_t w_\epsilon\|^2 + o_\epsilon(1)$$
$$= \tilde{E}_\epsilon(s_0,\psi_0,u_0) - \tilde{E}_\epsilon(w_\epsilon)(T)$$

for $T > 0$. By (4.11) and [15, pp. 427–437], we may obtain

$$\tilde{E}_\epsilon(w_\epsilon)(T) = \pi\,(d_1+d_2)\,\log\frac{1}{\epsilon} + W(\vec{a}(T)) + (d_1+d_2)\,\gamma - 2N_0\,\tilde{\gamma} + \rho(T) \tag{4.27}$$

for $T > 0$, where $\rho(T)$ comes from the weak convergence in H^1 of $w_\epsilon(\cdot,T)$. Note that $\rho(T) = 0$ if $w_\epsilon(\cdot,T)$ has strong convergence. By (4.21), we have

$$\tilde{E}_\epsilon(w_\epsilon)(0) = \pi\,(d_1+d_2)\,\log\frac{1}{\epsilon} + W(\vec{a}(0)) + (d_1+d_2)\,\gamma - 2N_0\,\tilde{\gamma} + o_\epsilon(1)\,. \tag{4.28}$$

Hereafter, $\vec{a}(0) = (a_1^0,\dots,a_{N_0}^0,b_1^0,\dots,b_{d_1-N_0}^0,c_1^0,\dots,c_{d_2-N_0}^0)$. Thus by (4.26)–(4.28), we obtain

$$\rho(T) \le W(\vec{a}(0)) - W(\vec{a}(T)) - \int_0^T \sum_{j=1}^{N_0} 2|\dot{a}_j|^2 + \sum_{k=1}^{d_1-N_0} |\dot{b}_k|^2 + \sum_{l=1}^{d_2-N_0} |\dot{c}_l|^2\,. \tag{4.29}$$

Since we do not know the strong convergence theorem for w_ϵ, the argument of Theorem 2.1(iii) in [15] can only ensure $\vec{a}(t)$ satisfying

$$M_0\,\frac{d}{d\,t}\,\vec{a}(t) = -\nabla W(\vec{a}(t)) + \zeta(t)\,, \tag{4.30}$$

where ζ is from the defect measure of weak convergence and $M_0 = \text{ diag }(M_0^i | i = 1,\dots,d_1+d_2)$ is a diagonal matrix such that $M_0^i = 2$ if $i = 1,\dots,N_0$ and $M_0^i = 1$ otherwise. By a simple energy comparison (cf. [16]), it is easy to check that

$$|\zeta(T)| \le C\,\rho(T) \quad \text{ for } T > 0\,, \tag{4.31}$$

where C is a positive constant independent of ϵ and T. Now we want to claim $\zeta \equiv 0$ and (4.14) is valid. Let $\vec{b}(t)$ be the solution of

$$M_0\,\frac{d}{d\,t}\,\vec{b}(t) = -\nabla W(\vec{b}(t))\,, \tag{4.32}$$

with initial data $\vec{b}(0) = \vec{a}(0)$. Then it is obvious that

$$W(\vec{a}(0)) = W(\vec{b}(T)) + \int_0^T \frac{d}{d\,t}\vec{b}(t)\cdot\left(M_0\,\frac{d}{d\,t}\vec{b}(t)\right)\,dt\,. \tag{4.33}$$

Hence, by (4.29) and (4.33), we have

$$\rho(T) \le C_0\,|\vec{a}(T) - \vec{b}(T)| + C_1\int_0^T \left|\frac{d}{d\,t}\,(\vec{a}-\vec{b})\right|(t)\,dt \tag{4.34}$$

for $T > 0$. Hereafter we set C_j's as positive constants independent of ϵ and T. From (4.30) and (4.32),

$$\left|\frac{d}{dt}(\vec{a}-\vec{b})\right|(T) \le C_2|\vec{a}-\vec{b}|(T) + |\zeta(T)|\,. \tag{4.35}$$

Thus, by (4.31), (4.34), and (4.35), we obtain

$$\left|\frac{d}{dt}(\vec{a}-\vec{b})\right|(T) \le C_3|\vec{a}-\vec{b}|(T) + C_4\int_0^T \left|\frac{d}{dt}(\vec{a}-\vec{b})\right|(t)\,dt \tag{4.36}$$

for $T > 0$. Therefore, by the Gronwall's inequality, we obtain $\vec{a} \equiv \vec{b}$; i.e., $\zeta \equiv 0$, and we complete the proof of Theorem 4.2.

Final remark. For the approximated gradient flow with the Neumann boundary condition, we may consider

(4.37)
$$\begin{cases} \partial_t w = -\text{ grad }\tilde{E}_\epsilon(w), & w(x,t) = (s,\psi,u)(x,t) \in \mathbb{C}_\delta & \text{for } x \in \Omega, t > 0, \\ \frac{\partial}{\partial \nu}(s,\psi,u)(x,t) = 0 & & \text{for } x \in \partial\Omega\,, t > 0, \\ s = s_0\,, \psi = \psi_0\,, & u = u_0 & \text{for } x \in \Omega\,, t = 0, \end{cases}$$

where $s_0 = (|\psi_0|^2 + \frac{1}{2}\delta^2)^{\frac{1}{2}}$ and $\frac{\partial}{\partial \nu}$ is the normal derivative on $\partial\Omega$. Here $(s_0,\psi_0,u_0)(x) \in \mathbb{C}_\delta$ for $x \in \Omega$ and

$$\begin{aligned}\tilde{E}_\epsilon(s_0,\psi_0,u_0) = \pi\,(d_1+d_2)\,\log\frac{1}{\epsilon} + W_1(a_1^0,\dots,a_{N_0}^0,b_1^0,\dots,b_{d_1-N_0}^0) \\ + W_2(a_1^0,\dots,a_{N_0}^0,c_1^0,\dots,c_{d_2-N_0}^0) + (d_1+d_2)\,\gamma - 2N_0\,\tilde{\gamma} + o_\epsilon(1),\end{aligned} \tag{4.38}$$

where a_j^0's and b_k^0's are distinct essential zeros of ψ_0, and a_j^0's and c_l^0's are distinct essential zeros of u_0, respectively. Hereafter, W_i's are defined by

$$\begin{aligned} &W_1(a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0}) \\ &\quad = \lim_{r\to 0+}\left\{\frac{1}{2}\int_{\Omega\setminus[\cup_{j=1}^{N_0} B_r(a_j)\ \cup_{k=1}^{d_1-N_0}\ B_r(b_k)]} |\nabla\Psi_{a,b}|^2\,dx - \pi\,d_1\,\log\frac{1}{r}\right\}, \\ &W_2(a_1,\dots,a_{N_0},c_1,\dots,c_{d_2-N_0}) \\ &\quad = \lim_{r\to 0+}\left\{\frac{1}{2}\int_{\Omega\setminus[\cup_{j=1}^{N_0} B_r(a_j)\ \cup_{k=1}^{d_2-N_0}\ B_r(c_l)]} |\nabla U_{a,c}|^2\,dx - \pi\,d_2\,\log\frac{1}{r}\right\}, \end{aligned}$$

where

$$\begin{aligned}\Psi_{a,b}(x) &= \prod_{j=1}^{N_0}\frac{x-a_j}{|x-a_j|}\prod_{k=1}^{d_1-N_0}\frac{x-b_k}{|x-b_k|}\,e^{i\,h_{a,b}(x)}\,, \\ U_{a,c}(x) &= \prod_{j=1}^{N_0}\frac{x-a_j}{|x-a_j|}\prod_{k=1}^{d_2-N_0}\frac{x-c_k}{|x-c_k|}\,e^{i\,h_{a,c}(x)}\,,\end{aligned}$$

and $h_{a,b}$ and $h_{a,c}$ are harmonic functions on Ω such that the value of $h_{a,b}$ and $h_{a,c}$ on $\partial\Omega$ is determined by the requirement $\frac{\partial}{\partial\nu}\Psi_{a,b} = \frac{\partial}{\partial\nu}U_{a,c} = 0$ on $\partial\Omega$. Then, by the

argument of Theorem 2.1 in [15] and the proof of Theorem 4.2, we may derive the dynamics of vortices as follows:

$$\begin{cases} -2\,\dot{a}_j &= \nabla_{a_j} W_1(a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0}) + \nabla_{a_j} W_2(a_1,\dots,a_{N_0},c_1,\dots,c_{d_2-N_0}),\\ -\dot{b}_k &= \nabla_{b_k} W_1(a_1,\dots,a_{N_0},b_1,\dots,b_{d_1-N_0}),\\ -\dot{c}_l &= \nabla_{c_l} W_2(a_1,\dots,a_{N_0},c_1,\dots,c_{d_2-N_0}). \end{cases}$$

for $j = 1,\dots,N_0, k = 1,\dots,d_1 - N_0$, and $l = 1,\dots,d_2 - N_0$.

Acknowledgment. The second author wishes to express his sincere thanks to B. Rosenstein for helpful discussions.

REFERENCES

[1] F. Bethuel, H. Brezis, and F. Helein, *Ginzburg-Landau Vortices*, Birkhäuser Boston, Boston, 1994.

[2] D. Bishop, C. Varma, B. Batlogg, E. Bucher, Z. Fisk, and J. Smith, *Ultrasonic attenuation in UPt_3*, Phys. Rev. Lett., 53 (1984), pp. 1009–1012.

[3] X. Chen, C. M. Elliott, and T. Qi, *Shooting method for vortex solutions of a complex-valued Ginzburg-Landau equation*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 1075–1088.

[4] J. E. Colliander and R. L. Jerrard, *Vortex dynamics for the Ginzburg-Landau-Schrödinger equation*, Internat. Math. Res. Notices, 7 (1998), pp. 333–358.

[5] W. Y. Ding and F. H. Lin, *A generalization of Eells-Sampson's theorem*, J. Partial Differential Equations, 5 (1992), pp. 13–22.

[6] J. L. Ericksen, *Liquid crystals with variable degree of orientation*, Arch. Ration. Mech. Anal., 113 (1990), pp. 97–120.

[7] R. Fisher, S. Kim, B. Woodfield, N. Philips, N. Taillefer, K. Hasselback, J. Floquet, A. Giorgi, and J. Smith, *Specific heat of UPt_3: Evidence for unconventional superconductivity*, Phys. Rev. Lett., 62 (1989), pp. 1411–1414.

[8] P. G. de Gennes, *Superconductivity of Metals and Alloys*, Addison–Wesley, Reading, MA, 1989.

[9] P. S. Hagan, *Spiral waves in reaction-diffusion equations*, SIAM J. Appl. Math., 42 (1982), pp. 762–786.

[10] R. M. Hervé and M. Hervé, *Étude qualitative des solutions réelles d'une équation différentielle liée à l'équation de Ginzburg-Landau*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 11 (1994), pp. 427–440.

[11] A. Knigavko and B. Rosenstein, *Spontaneous vortex state and ferromagnetic behavior of type-*II *p-wave superconductors*, Phys. Rev. B, 58 (1998), pp. 9354–9364.

[12] A. J. Leggett and S. Takagi, *NMR in $A-^3He$ and $B-^3He$: The intrinsic relaxation mechanism*, Phys. Rev. Lett., 34 (1975), pp. 1424–1427.

[13] F. H. Lin, *On nematic liquid crystals with variable degree of orientation*, Comm. Pure Appl. Math., 44 (1991), pp. 453–468.

[14] F. H. Lin, *Static and moving vortices in Ginzburg-Landau theories*, in Progr. Nonlinear Differential Equations Appl. 29, Birkhäuser Verlag, Basel, 1997, pp. 71–111.

[15] F. H. Lin, *Complex Ginzburg-Landau equations and dynamics of vortices, filaments, and codimension-*2 *submanifolds*, Comm. Pure Appl. Math., 51 (1998), pp. 385–441.

[16] F. H. Lin, *Vortex dynamics for the nonlinear wave equation*, Comm. Pure Appl. Math., 52 (1999), pp. 737–761.

[17] F. H. Lin and T. C. Lin, *Minimax solutions of the Ginzburg-Landau equations*, Selecta Math. (N.S.), 3 (1997), pp. 99–113.

[18] F. H. Lin and J. X. Xin, *On the incompressible fluid limit and the vortex motion law of the nonlinear Schrödinger equation*, Comm. Math. Phys., 200 (1999), pp. 249–274.

[19] Y. Maeno, H. Hashimoto, K. Yoshida, S. Nishizaki, T. Fujita, J. G. Bednortz, and F. Lichtenberg, *Superconductivity in a layered perovskite without copper*, Nature, 372 (1994), pp. 532–534.

[20] F. Pacard and T. Riviére, *Linear and Nonlinear Aspects of Vortices. The Ginzburg-Landau Model*, Birkhäuser Boston, Boston, 2000.

[21] E. Sandier, *Lower bounds for the energy of unit vector fields and applications*, J. Funct. Anal., 152 (1998), pp. 379–403.

[22] B. Shivaram, Y. Jeong, T. Rosenbaum, and D. Hinks, *Anisotropy of transverse sound in the heavy fermion superconductor UPt_3*, Phys. Rev. Lett., 56 (1986), pp. 1078–1081.

[23] B. Shivaram, T. Rosenbaum, and D. Hinks, *Unusual angular and temperature dependence of the upper critical field in UPt_3*, Phys. Rev. Lett., 57 (1986), pp. 1259–1262.

[24] M. Sigrist and K. Ueda, *Phenomenological theory of conventional superconductivity*, Rev. Modern Phys., 63 (1991), pp. 239–311.

[25] P. J. C. Signore, J. P. Koster, E. A. Knetsch, C. M. V. Woerkens, M. W. Meisel, S. E. Brown, and Z. Fisk, *Inductive response of oriented UPt_3 in the superconducting state*, Phys. Rev. B, 45 (1992), pp. 10151–10154.

[26] M. Struwe, *On the asymptotic behavior of minimizers of the Ginzburg-Landau model in* 2 *dimensions*, Differential Integral Equations, 7 (1994), pp. 1613–1624.

[27] M. Struwe, *Erratum: "On the asymptotic behavior of minimizers of the Ginzburg-Landau model in* 2 *dimensions,"* Differential Integral Equations, 8 (1995), p. 224.

[28] H. Tou, Y. Kitaoka, K. Asayama, K. Kimura, Y. Onuki, E. Yamamoto, and K. Maezawa, *Odd-parity superconductivity with parallel spin pairing in UPt_3: Evidence from ^{195}Pt Knight shift study*, Phys. Rev. Lett., 77 (1996), pp. 1374–1377.

[29] H. Tou, Y. Kitaoka, K. Asayama, K. Kimura, Y. Onuki, E. Yamamoto, and K. Maezawa, *Nonunitary spin-triplet superconductivity in UPt_3: Evidence from ^{195}Pt Knight shift study*, Phys. Rev. Lett., 80 (1998), pp. 3129–3132.

[30] J. X. Zhu, C. S. Ting, J. L. Shen, and Z. D. Wang, *Ginzburg-Landau equations for layered p-wave superconductors*, Phys. Rev. B, 56, (1997), pp. 14093–14101.

SIAM J. MATH. ANAL.
Vol. 34, No. 5, pp. 1128–1151

STUDY OF THE BUCKLING OF A TAPERED ROD WITH THE GENUS OF A SET*

GRÉGORY VUILLAUME†

Abstract. This paper, which can be considered a continuation of the papers [C. A. Stuart, *J. Math. Pures Appl.*, 80 (2001), pp. 281–337] and [C. A. Stuart, *Proc. Roy. Soc. Edinburgh Sect. A*, 132 (2002), pp. 729–764], is concerned with the study of the buckling of a tapered rod. This physical phenomenon leads to the nonlinear eigenvalue problem

$$\{A(s)u'(s)\}' + \mu \sin u(s) = 0 \quad \text{for all } s \in (0,1),$$
$$u(1) = \lim_{s\to 0} A(s)u'(s) = 0,$$
$$\int_0^1 A(s)u'(s)^2 ds < \infty,$$

where $A(s) \in C([0,1])$ is such that $A(s) > 0$ for all $s > 0$ and $\lim_{s\to 0} A(s)/s^p = L$ for some constants $p \geq 0$ and $L \in (0,\infty)$. We study the set of all solutions of the problem and, in particular, find the points $\mu \in \mathbb{R}_+$ such that bifurcation occurs at $(\mu, 0)$.

As was shown by Stuart in [*J. Math. Pures Appl.*, 80 (2001), pp. 281–337], there is a number $\Lambda(A) \geq 0$ such that, for $\mu \leq \Lambda(A)$, $u \equiv 0$ is the only solution of the problem, and it minimizes the energy in the space of all admissible configurations. For $\mu > \Lambda(A)$, the energy is minimized by a nontrivial solution. For $0 \leq p < 2$, bifurcation occurs at a discrete set of eigenvalues μ_i, $i \in \mathbb{N}^* = \{1, 2, \dots\}$, which satisfy $\mu_1 = \Lambda(A)$, $\mu_i < \mu_{i+1}$ for all $i \in \mathbb{N}^*$ and $\lim_{i\to\infty} \mu_i = \infty$. At $p = 2$, changes occur. For $0 \leq p \leq 2$, $\Lambda(A) > 0$, whereas $\Lambda(A) = 0$ for $p > 2$. For $p = 2$, there is a number $\Lambda_e(A) \in [\Lambda(A), \infty)$ such that bifurcation occurs at every value $\mu \in [\Lambda_e(A), \infty)$. In this paper, we show the following points:

- For $p = 2$, if $\Lambda(A) < \Lambda_e(A)$, bifurcation from the solution $u \equiv 0$ also occurs at a finite or countable set of eigenvalues $\mu_i \in I \subset \mathbb{N}^*$, where $\mu_1 = \Lambda(A)$ and $\mu_i < \Lambda_e(A)$ for all $i \in I$.
- For $2 < p < 3$, bifurcation occurs at every value $\mu \geq 0$.

Key words. nonlinear eigenvalue problem, bifurcation, genus of a set

AMS subject classification. 47J10

PII. S0036141002404322

1. Introduction.

We begin by recalling the definition of the problem we consider in this paper. For this, the following definition is crucial.

DEFINITION 1.1. *A profile for a column with tapering of order $p \geq 0$ is a function $A \in C([0,1])$ such that $A(s) > 0$ for $0 < s \leq 1$, and there exists $L \in (0,\infty)$ such that* $\lim_{s\to 0} \frac{A(s)}{s^p} = L$.

For such a profile there exist constants $K_1 \geq K_2 > 0$ such that

$$K_2 s^p \leq A(s) \leq K_1 s^p \quad \text{for all } s \in [0,1]. \tag{1}$$

We now give the formal statement of the mathematical problem to be considered. We recall that this problem represents one of the simplest models for the planar buckling of a tapered column. For more details, see [8]. Consider a profile A with tapering of order $p \geq 0$ and a constant $\mu \geq 0$.

*Received by the editors March 22, 2002; accepted for publication (in revised form) November 9, 2002; published electronically April 15, 2003.

http://www.siam.org/journals/sima/34-5/40432.html

†Institut de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland (gregory.vuillaume@epfl.ch).

DEFINITION 1.2. *A solution of problem P is a function* $u \in C^1((0,1])$ *such that* $Au' \in C^1((0,1])$,

$$\{A(s)u'(s)\}' + \mu \sin u(s) = 0 \quad \text{for all } s \in (0,1], \tag{2}$$

$$u(1) = \lim_{s \to 0} A(s)u'(s) = 0, \tag{3}$$

$$\int_0^1 A(s)u'(s)^2 ds < \infty. \tag{4}$$

In fact, for a given profile, we would like to find the parameters μ where bifurcation from the trivial solution $u \equiv 0$ occurs. Clearly problem P is a nonlinear eigenvalue problem of Sturm–Liouville type, but it becomes increasingly singular at $s = 0$ as the order of tapering p gets bigger. For the cases $p \geq 2$, which we deal with here, the singularity is so severe that the problem cannot be treated by standard tools from bifurcation theory such as the Crandall–Rabinowitz theorem. Indeed, for $p > 2$, the problem does not have a rigorous linearization about the trivial solution $u \equiv 0$.

Stuart (see [8], [9]) showed that in several respects (the shape of the buckled configurations, the nature of the bifurcation diagrams) tapering of order 2 plays a critical role, in the sense that the situation when $p < 2$ is very different from what happens when $p \geq 2$. He proved that there is a number $\Lambda(A) \geq 0$ such that, for $\mu \leq \Lambda(A)$, $u \equiv 0$ is the only solution of the problem, and it minimizes the energy in the space of all admissible configurations that will be denoted by H_A. For $\mu > \Lambda(A)$, the energy is minimized by a nontrivial solution. For $0 \leq p < 2$, bifurcation occurs at a discrete set of eigenvalues μ_i, $i \in \mathbb{N}^* = \{1, 2, \dots\}$, which satisfy $\mu_1 = \Lambda(A)$, $\mu_i < \mu_{i+1}$ for all $i \in \mathbb{N}^*$, and $\lim_{i \to \infty} \mu_i = \infty$. At $p = 2$, changes occur. For $0 \leq p \leq 2$, $\Lambda(A) > 0$, whereas $\Lambda(A) = 0$ for $p > 2$. For $p = 2$, there is a number $\Lambda_e(A) \in [\Lambda(A), \infty)$ such that bifurcation occurs at every value $\mu \in [\Lambda_e(A), \infty)$. In this paper, we show the following points:

- For $p = 2$, if $\Lambda(A) < \Lambda_e(A)$, bifurcation from the solution $u \equiv 0$ occurs also at a finite or countable set of eigenvalues $\mu_i \in I \subset \mathbb{N}^*$, where $\mu_1 = \Lambda(A)$ and $\mu_i < \Lambda_e(A)$ for all $i \in I$.
- For $2 < p < 3$, bifurcation occurs at every value $\mu \geq 0$.

Now we introduce the Hilbert space of all admissible configurations.

1.1. The energy space H_A. Consider $p \in [0, \infty)$. If an element $u \in L^1_{loc}((0,1])$ admits a generalized derivative u' on $(0,1)$ such that $\int_0^1 s^p u'(s)^2 ds < \infty$, it follows that $u \in W^{1,1}((\epsilon, 1))$ for all $\epsilon \in (0,1)$ and hence, from Theorem VIII.2 of [1], that (after a modification on a set of measure zero) $u \in C((0,1])$. For $p \geq 0$, let

$$H_p = \left\{ u \in L^1_{loc}((0,1]) : \int_0^1 s^p u'(s)^2 ds < \infty \text{ and } u(1) = 0 \right\}.$$

The next proposition recalls some important properties concerning the space H_p (for the proof; see [8]).

PROPOSITION 1.3.

1. *For* $p \in [0, \infty)$, H_p *with the scalar product*

$$\langle u, v \rangle_p = \int_0^1 s^p u'(s) v'(s) ds$$

is a Hilbert space.

2. *For any bounded sequence* $\{u_n\}$ *in* H_p *there exist a function* $u \in C((0,1])$ *and a subsequence* $\{u_{n_k}\}$ *such that* $u_{n_k} \to u$ *uniformly on* $[\epsilon, 1]$ *for every* $\epsilon \in (0,1)$.
3. $H_p \cap L^\infty(0,1)$ *is dense in* H_p.
4. *If* $u \in H_p$, *then so does* $|u|$, *and* $|u|'(s)^2 = u'(s)^2$ *almost everywhere (a.e.) on* $(0,1)$.

Remark. If A is a profile for a column with tapering of order p, then

$$\langle u, v\rangle_A = \int_0^1 A(s)u'(s)v'(s)ds$$

is a scalar product on $H_A = H_p$ whose norm is equivalent to $\|.\|_p$. Indeed, we have

$$\sqrt{K_2}\|u\|_p \le \|u\|_A \le \sqrt{K_1}\|u\|_p \quad \text{for all } u \in H_A, \tag{5}$$

with the constants given in (1). The Hilbert space $(H_A, \langle .,.\rangle_A)$ will be referred to as the *energy space for the profile A*. If the sequence $\{u_n\}$ converges weakly to u in H_A, then $u_n \to u$ uniformly on $[\epsilon, 1]$ for every $\epsilon \in (0,1)$.

Now we are able to state our first main result.

THEOREM 1.4. *Let A be a profile with tapering of order* $2 < p < 3$ *and consider* $\mu > 0$. *For this value of* μ, *there are infinitely many solutions* $\{u_k\}$ *of problem P with the property that* $|u_k(s)| < \pi$ *for all* $s \in (0,1]$. *Furthermore,* $\|u_k\|_A \to 0$ *as* $k \to 0$ *and the number of zeros of* u_k *tends to infinity as* $k \to \infty$.

A proof of this result is given in section 4, after we give some preliminaries in the rest of this section and in sections 2 and 3.

1.2. The linearized problem. To understand our next main results, we need to introduce the "linearization" of problem P.

DEFINITION 1.5. *A solution of problem PL is a function* $u \in C^1((0,1])$ *such that* $Au' \in C^1((0,1])$,

$$\{A(s)u'(s)\}' + \mu u(s) = 0 \quad \text{for all } s \in (0,1], \tag{6}$$

and (3) *and* (4) *are satisfied. If* $u \neq 0$, *it is called an eigenfunction associated with the eigenvalue* μ.

Now we consider profiles with tapering of order $p \in [0,2]$ and summarize the results obtained by Stuart in [9] concerning problem PL. We first introduce a bounded linear operator $T : H_A \to H_A$ associated with this problem. All proofs of these results can be found in [9].

PROPOSITION 1.6. *Let A be a profile with tapering of order* $p \in [0,2]$. *There is a unique bounded linear operator* $T : H_A \to H_A$ *such that*

$$\langle Tu, v\rangle_A = \int_0^1 u(s)v(s)ds \quad \text{for all } u, v \in H_A. \tag{7}$$

Furthermore T is a positive self-adjoint operator in H_A *and* 0 *is not an eigenvalue of T. For* $p < 2$, $T : H_A \to H_A$ *is also compact.*

The spectrum of T, the discrete spectrum of T, and the essential spectrum of T are the sets defined, respectively, by

$$\begin{aligned}
\sigma(T) &= \{\lambda \in \mathbb{R} : T - \lambda I : H_A \to H_A \text{ is not an isomorphism}\},\\
\sigma_d(T) &= \{\lambda \in \sigma(T) : T - \lambda I : H_A \to H_A \text{ is a Fredholm operator}\},\\
\sigma_e(T) &= \sigma(T) \setminus \sigma_d(T).
\end{aligned}$$

Note that $\sigma_d(T)$ is formed by the isolated eigenvalues of T which have finite multiplicity (see Theorem 1.6 in Chapter IX of [3]). Since T is positive and self-adjoint, we know that $\sigma(T) \subset [0, \infty)$ and

$$\begin{aligned}\|T\| = \max \sigma(T) &= \sup \Big\{ \langle Tu, u\rangle_A : u \in H_A \text{ with } \|u\|_A = 1 \Big\} \\ &= \sup \left\{ \frac{\langle Tu, u\rangle_A}{\langle u, u\rangle_A} : u \in H_A \setminus \{0\} \right\}.\end{aligned}$$

We can express this by defining the Rayleigh quotient,

$$Q_A(u) = \frac{\int_0^1 A(s)u'(s)^2 ds}{\int_0^1 u(s)^2 ds} \tag{8}$$

(we set $Q_A(u) = 0$ if $\int_0^1 u(s)^2 ds = \infty$) and its infimum

$$\Lambda(A) = \inf \big\{ Q_A(u) : u \in H_A \setminus \{0\} \big\}. \tag{9}$$

For $p > 2$, using the functions u_α defined below in (16) with $(1-p)/2 < \alpha < -1/2$, we have that

$$0 < \int_0^1 A(s)u'_\alpha(s)^2 ds < \infty \quad \text{and} \quad \int_0^1 u_\alpha(s)^2 ds = \infty.$$

This implies that

$$\Lambda(A) = 0 \quad \text{if } p > 2. \tag{10}$$

Now if $p \in [0, 2]$, it follows from Lemma 2.1 and (5) that

$$\Lambda(A) \geq \frac{K_2}{4} > 0. \tag{11}$$

In this case, we have

$$\|T\| = \max \sigma(T) = 1/\Lambda(A), \tag{12}$$

and $\Lambda(A)$ is the infimum of the spectrum of problem PL since the eigenfunctions of problem PL are precisely the eigenfunctions of the operator T, as is shown by the next proposition.

PROPOSITION 1.7. *Let A be a profile with tapering of order $p \in [0, 2]$. Then u is an eigenfunction of problem PL if and only if $u \in H_A \setminus \{0\}$ and $u = \mu Tu$. Furthermore, all eigenvalues of T are simple.*

THEOREM 1.8. *Let A be a profile with tapering of order $0 \leq p < 2$. Then*

$$\sigma_d(T) = \big\{ \lambda_i : i \in \mathbb{N}^* \big\} \quad \textit{and} \quad \sigma_e(T) = \{0\},$$

where $\lambda_{i+1} < \lambda_i$, $\lambda_1 = \Lambda(A)^{-1}$, $\lim_{i\to\infty} \lambda_i = 0$, and each λ_i is a simple eigenvalue of T.

The preceding theorem shows that in the case $0 \leq p < 2$ problem PL behaves like a regular Sturm–Liouville problem, in particular $\sigma_e(T) = \{0\}$. For $p = 2$, the situation changes. We always have $\max \sigma_e(T) > 0$, and we may have $\sigma_d(T) = \emptyset$. That is what proves the following results.

THEOREM 1.9. *Let A be a profile with tapering of order $p = 2$. Then $\max \sigma_e(T) = 4/L$, where $L = \lim_{s\to 0} A(s)/s^2$ and $T : H_A \to H_A$ is not compact.*

Note that $\sigma(A) = \{\mu = 1/\lambda : \lambda \in \sigma(T) \setminus \{0\}\}$, so $\Lambda(A) = \inf \sigma(A)$ is the infimum of the spectrum of problem PL. We introduce the following notation for the infimum of the essential spectrum of problem PL:

$$\Lambda_e(A) = \inf \sigma_e(A),$$

where $\sigma_e(A) = \{\mu = 1/\lambda : \lambda \in \sigma_e(T) \setminus \{0\}\}$.

THEOREM 1.10. *Let A be a profile with tapering of order $p = 2$. Then*

$$0 < \frac{K_2}{4} \le \Lambda(A) \le \frac{L}{4} = \Lambda_e(A),$$

where $L = \lim_{s\to 0} A(s)/s^2$ and $K_2 = \inf_{0<s\le 1} A(s)/s^2$.

In particular, $\Lambda(A) = \Lambda_e(A) = L/4$, provided that

$$A(s) \ge Ls^2 \quad \textit{for all } s \in (0, 1].$$

Remark 1.1. To show that $\Lambda(A) < \Lambda_e(A)$ it is sufficient to find one function $u \in H_2$ such that $Q_A(u) < L/4$. As is shown in [9], this can be done, provided that

$$\frac{\pi^2 \max_{s\in I} A(s)}{|I|\{2\delta + |I|\}} < L$$

for some interval $I = [\delta, \gamma] \subset (0, 1]$.

Remark 1.2. In some cases, there may be no eigenfunctions at all. For example, if $A(s) = Ls^2$ for all $s \in [0, 1]$, then T has no eigenvalues and $u \equiv 0$ is the only solution of problem PL.

Remark 1.3. We have been able to prove that for each $N \in \mathbb{N}^*$, there exists a profile $A(s)$ with tapering of order 2 such that $T : H_A \to H_A$ has at least N simple characteristic values $\lambda_1 = \Lambda(A) < \lambda_2 < \cdots < \lambda_N < \Lambda_e(A)$.

We now are able to express our next main results. We make the following assumption:

$$\text{(H)} \qquad \begin{cases} A(s) \text{ is a profile with tapering of order 2 such that} \\ \\ \qquad\qquad \Lambda(A) < \Lambda_e(A). \end{cases}$$

Notation. Under the assumption (H), the operator $T : H_A \to H_A$ defined by (7) has at least one simple eigenvalue $\Lambda(A)^{-1}$. In fact, T has a finite or countable number of simple eigenvalues $\{\mu_i^{-1} : i \in I \subset \mathbb{N}^*\}$ such that $\mu_i^{-1} > \mu_{i+1}^{-1} > \max \sigma_e(T)$ for all $i \in I$ and $\sigma_d(T) \cap (\Lambda_e(A)^{-1}, \Lambda(A)^{-1}] = \{\mu_i^{-1} : i \in I\}$. We have $\mu_1 = \Lambda(A)$ and $\mu_i < \Lambda_e(A)$ for each $i \in I$. For $i \in I$, we note $\varphi_i \in H_A$, the eigenvector associated to μ_i such that $\|\varphi_i\|_A = 1$. We then have $\varphi_i = \mu_i T\varphi_i$ for all $i \in I$.

For each $i \in I$, we note μ_i^+ for $\min\{\mu \in \sigma(A) : \mu > \mu_i\}$. In our context, we have that $\mu_i^+ = \mu_{i+1}$ if T has at least $(i+1)$ eigenvalues above $\max \sigma_e(T)$ and $\mu_i^+ = \Lambda_e(A)$ if T has exactly i eigenvalues above $\max \sigma_e(T)$.

THEOREM 1.11. *Under assumption* (H)*, and with the corresponding notation, choose $i \in I$. Consider $\mu > \mu_i$. For this value μ, problem P has at least i nontrivial solutions $\{u_k\}$, $k = 1, \ldots, i$, with the property that $|u_k(s)| < \pi$ and $u_k'(1) < 0$ for all $s \in (0, 1]$ and for all $k = 1, \ldots, i$.*

Remark. In fact, in the context of Theorem 1.11, we could show that problem P admits at least $2i$ solutions, since if u is a solution, then $-u$ is another solution.

Moreover, we have the following bifurcation result.

THEOREM 1.12. *Under assumption* (H), *and with the corresponding notation, choose* $i \in I$. *Then bifurcation from the trivial solution* $u \equiv 0$ *occurs at* $(\mu_i, 0) \in \mathbb{R} \times H_A$ *in the sense that there exists a sequence* $\{(\mu^n, u_n)\}_{n\geq 1} \in \mathbb{R} \times H_A \setminus \{0\}$ *such that* (μ^n, u_n) *is a solution of problem P for all* $n \geq 1$, $\lim_{n\to\infty} \mu^n = \mu_i$, *and* $\lim_{n\to\infty} \|u_n\|_A = 0$.

To have a better comprehension of the way we prove these results, we now give some explanations of the tools used by Stuart to prove his results in [8]. After that, we will explain the main differences between this work and that of Stuart.

For a profile with tapering of any order $p \geq 0$ and a constant $\mu > 0$, Stuart defined in [8] the following energy functional:

$$J_\mu(u) = \frac{1}{2}\|u\|_A^2 - \mu \int_0^1 \{1 - \cos u(s)\} ds. \tag{13}$$

For $p > 2$, J_μ may not be Fréchet differentiable. However, for all $p \geq 0$ and $u \in H_A$, we have that

$$\frac{d}{dt} J_\mu(u + tv)|_{t=0} = \int_0^1 A(s)u'(s)v'(s)ds - \mu \int_0^1 v(s) \sin u(s) ds$$

for all $v \in H_A \cap L^1(0,1)$. We recall from Proposition 1.3 that $H_A \cap L^\infty(0,1)$ is dense in H_A. Thus J_μ has directional derivatives at u for all directions in a dense subspace of H_A. The solutions of problem P are related to stationary points of J_μ in the following way (a proof of all these facts can be found in [8]).

THEOREM 1.13. *Let* A *be a profile with tapering of order* $p \geq 0$.

(i) *A function* u *is a solution of problem P if and only if* $u \in H_A$ *and*

$$\int_0^1 A(s)u'(s)v'(s)ds = \mu \int_0^1 v(s) \sin u(s) ds$$

for all $v \in H_A \cap L^1(0,1)$.

(ii) *For* $p \in [0,2]$, $J_\mu \in C^1(H_A)$ *and a function* u *is a solution of problem P if and only if* $u \in H_A$ *and* $J'_\mu(u) = 0$. *Moreover, there exists a completely continuous function* $G_A : H_A \to H_A$ *such that*

$$J'_\mu(u)v = \langle u - \mu G_A(u), v\rangle_A \quad \text{for all } u, v \in H_A.$$

If A is a profile with tapering of order $p \in [0,2]$ and if $F : \mathbb{R} \times H_A \to H_A$ is defined by

$$F(\mu, u) = u - \mu G_A(u),$$

we have by Theorem 1.13 that (μ, u) is a solution of problem P if and only if $u \in H_A$ and $F(\mu, u) = 0$.

In the case $0 \leq p < 2$, F is continuously Fréchet differentiable with $DF(\mu, 0) = I - \mu T$, where I denotes the identity mapping in H_A and T is the positive self-adjoint and compact operator on H_A defined by (7). In this case, Theorem 1.8 tells us that $\sigma(T) = \sigma_d(T) \cup \sigma_e(T)$, where $\sigma_e(T) = \{0\}$ and $\sigma_d(T) = \{\mu_i^{-1} : i \in \mathbb{N}^*\}$ with $\mu_{i+1}^{-1} < \mu_i^{-1}$, $\mu_1^{-1} = \Lambda(A)^{-1}$, $\lim_{i\to\infty} \mu_i^{-1} = 0$ and each μ_i^{-1} is a simple eigenvalue of

T. Using standard results like the theorem of Crandall and Rabinowitz concerning bifurcation from simple eigenvalues or the theorem of Rabinowitz concerning global bifurcation (it is possible to use these results since F is Fréchet differentiable), it is proved in [8] that global bifurcation from the trivial solution $u \equiv 0$ occurs at the discrete set of characteristic values $\{\mu_i : i \in \mathbb{N}^*\}$ of T.

For $p = 2$, the situation is different. The function F is not Fréchet differentiable at $(\mu, 0)$ anymore, but only Gâteaux differentiable, and the Gâteaux derivative of F at $(\mu, 0)$ is still of the form $I - \mu T$, but now T is self-adjoint and not compact. Moreover, $\max \sigma_e(T) = \Lambda_e(A)^{-1} > 0$ (see Theorems 1.9 and 1.10). In this case, the results in [8] show that bifurcation occurs at $\mu = \Lambda(A)$ and also at every $\mu \in [\Lambda_e(A), \infty)$. For $p = 2$, the operator T may or may not have eigenvalues depending on the form of the profile A (see Remarks 1.1, 1.2, 1.3 following Theorem 1.10).

Remark. To see that F is not Fréchet differentiable at $(\mu, 0)$ for $p = 2$, we only have to show that G_A is not Fréchet differentiable at $u = 0$. To see this, suppose that G_A is Fréchet differentiable at $u = 0$. Since T is the Gâteaux derivative of G_A at 0, we have that $G'_A(0) = T$. But then T would be compact by Lemma 4.1 of [5, p. 135] and part (ii) of Theorem 1.13, which implies that G_A is compact. But this is a contradiction of Theorem 1.9.

In the present work, we suppose that A is a profile with tapering of order $2 \leq p < 3$. Since F is not Fréchet differentiable for $p = 2$ and is not even defined for $2 < p < 3$, we cannot use standard results like the ones mentioned above for the case $0 \leq p < 2$. However, for a profile A with tapering of order $p = 2$, introducing a modified version of the energy functional $J_\mu : H_A \to \mathbb{R}$, denoted by $j_\mu : H_A \to \mathbb{R}$, Stuart was able to show the following points (see [8] for the proofs):

(1) $j_\mu \in C^1(H_A)$ for all $\mu > 0$.
(2) If $\nabla j_\mu(u) = 0$ for some $\mu > 0$ and $u \in H_A$, then u is a solution of problem P and $|u(s)| < \pi$ for all $s \in (0, 1]$.
(3) j_μ satisfies the hypotheses of Theorem 2.5.

Then, applying Theorem 2.5 to the functional j_μ with $\mu > \Lambda_e(A)$, he obtained, for this value of μ, that there are infinitely many solutions $\{u_k\}$ of problem P and that $\|u_k\|_A \to \infty$ as $k \to \infty$. Then every $\mu \in [\Lambda_e(A), \infty)$ is a bifurcation point (see Theorem 5.8 of [8]).

We first consider profiles with tapering of order $2 < p < 3$. For such a profile, the functional j_μ can also be defined, and we are able to show that j_μ satisfies the hypotheses of Theorem 2.5 for all $\mu > 0$ (recall that in this case we have $\Lambda(A) = 0$). This is done in section 3, after some preliminaries are given in section 2. Section 4 is devoted to proving Theorem 1.4. Section 4 begins with the crucial Lemma 4.1 that will allow us to estimate the quantities $b_k(j_\mu)$ given in Theorem 2.5. For $p = 2$, Stuart obtained such estimates by using properties of the operator T. In our case, T is not defined and the main idea to prove Lemma 4.1 is to transform our profile A with tapering of order $2 < p < 3$ to a profile A_δ with tapering of order 2. Then, using Lemma 4.1, Theorem 1.4 can be proved in a similar way to Theorem 5.8 of [8].

In section 5, the approach is similar (that is, we apply Theorem 2.5 to the functional j_μ), but the situation and the conclusion are rather different. Here we consider a profile with tapering of order 2 satisfying assumption (H) with the corresponding notation. Stuart proved that for every $\mu > \Lambda_e(A)$, there are infinitely many solutions of problem P. Now, if we suppose that $\mu \in (\Lambda(A), \Lambda_e(A))$ and that $\mu_i < \mu$, then j_μ has at least i critical points, for in this case we are not able to prove that $b_k(j_\mu) < 0$ for infinitely many k, as it was the case for $\mu > \Lambda_e(A)$, or for $\mu > 0$ in the case of a

profile A with tapering of order $2 < p < 3$. In this case, we are able to prove only that $b_k(j_\mu) < 0$ for $k = 1, \dots, i$ (see the proof of Theorem 1.11). To prove that bifurcation occurs at μ_i (that is, to prove Theorem 1.12), we need the important Proposition 5.7, which states that $\lim_{\mu \to \mu_i} b_i(j_\mu) = 0$. See section 5.

After proving these results, a natural question arises. In the case of a profile with tapering of order 2 satisfying assumption (H), we have by Theorem 1.12 that bifurcation from the trivial solution occurs at each $(\mu, 0) \in \mathbb{R} \times H_A$, where μ is any characteristic value of T such that $\mu \in (\Lambda(A), \Lambda_e(A))$. Moreover, we know by Theorem 5.8 of [8] that bifurcation occurs at each $\mu \in [\Lambda_e(A), \infty)$. The question is, Do we have all bifurcation points, or is it possible that bifurcation occurs at a point $\mu \in (\Lambda(A), \Lambda_e(A))$ such that μ^{-1} does not belong to $\sigma(T)$ (recall that for $\mu < \Lambda(A)$, there is no nontrivial solution)? In fact, this is not a standard result because our problem, which can be written under the form $u = \mu G_A(u)$, is such that G_A is not Fréchet differentiable in 0. So we cannot conclude that every bifurcation value belongs to the spectrum of T^{-1} using standard results.

At first sight, this conclusion seems rather routine for a nonlinear eigenvalue like problem P. But we point out that we have an example of the following situation. H is a real Hilbert space and $G : H \to H$ is a Lipschitz continuous function such that $G(0) = 0$. Furthermore, G is Gâteaux differentiable at zero and $G'(0) : H \to H$ is a bounded self-adjoint operator. Nonetheless, the equation $u = \mu G(u)$ has bifurcation points $(\mu, 0)$ such that $\mu^{-1} \notin \sigma(G'(0))$. Of course G is not Fréchet differentiable at 0 since it is well known that $\mu^{-1} \in \sigma(G'(0))$ for all bifurcation points $(\mu, 0)$ in that case. Nevertheless, in our case, despite the lack of Fréchet differentiability, and making an additional assumption on the profile A, we have been able to prove that bifurcation cannot occur at a point $\mu \in (\Lambda(A), \Lambda_e(A))$ that is not a characteristic value of T. This is the subject of a forthcoming paper.

2. Preliminaries. In this part, we introduce the tools that we will need to prove our main results.

2.1. Some properties on H_A. We begin by recalling some properties of the Hilbert space H_A obtained by Stuart [8]. By Proposition 1.3, we have that

$$\|u\|_p = \left\{ \int_0^1 s^p u'(s)^2 ds \right\}^{1/2}$$

is a norm on the linear space H_p and we have the following estimation: For $u \in H_p$ and $s \in (0, 1]$,

$$|u(s)| \le \|u\|_p \left\{ \frac{1 - s^{1-p}}{1-p} \right\}^{1/2} \quad \text{if } p \neq 1, \tag{14}$$

whereas

$$|u(s)| \le \|u\|_p \left\{ \ln \frac{1}{s} \right\}^{1/2} \quad \text{if } p = 1. \tag{15}$$

Remark. Setting

$$u_\alpha(s) = s^\alpha (1 - s) \quad \text{for } 0 < s \le 1, \tag{16}$$

we see that $u_\alpha \in H_p \Leftrightarrow \alpha > (1-p)/2$. The following result shows that $H_p \subset L^2(0,1)$ for $0 \le p \le 2$. For $p > 2$ and $\alpha \in ((1-p)/2, -1/2]$, the function u_α defined by (16) belongs to H_p but not to $L^2(0,1)$.

Notation. In this work, the norm on $L^p(0,1)$ for $1 \le p \le \infty$ will be denoted by $\|.\|_{L^p}$ and the usual scalar product on $L^2(0,1)$ will be denoted by $\langle ., . \rangle_{L^2}$.

LEMMA 2.1. *Let $0 \le p \le 2$. Then $H_p \subset L^2(0,1)$ and*

$$\left\{ \int_0^1 u(s)^2 ds \right\}^{1/2} \le 2\|u\|_p \quad \textit{for all } u \in H_p. \tag{17}$$

Now, in particular, consider a profile A with tapering of order $2 < p < 3$. We recall that $\Lambda(A) = 0$. Using estimate (14) and inequality (5), we have that

$$\begin{aligned} \int_0^1 |u(s)| ds &\le \|u\|_p \int_0^1 \left\{ \frac{s^{1-p} - 1}{p-1} \right\}^{1/2} ds \\ &\le \frac{\|u\|_A}{\sqrt{K_2(p-1)}} \int_0^1 s^{(1-p)/2} ds < \infty \quad \text{for all } u \in H_A, \end{aligned}$$

since $s^{(1-p)/2}$ is integrable on $(0,1)$. Thus there exists a constant $C_1 > 0$ such that

$$\|u\|_{L^1} \le C_1 \|u\|_A \quad \text{for all } u \in H_A. \tag{18}$$

This implies that $H_A \subset L^1(0,1)$. In fact, we have a little bit more. That is what the next lemma shows.

LEMMA 2.2. *Let A be a profile with tapering of order $2 < p < 3$. Then there exist a number $\alpha(p) \in (1,2)$ and a constant $C > 0$ such that*

$$\|u\|_{L^{\alpha(p)}} \le C \|u\|_A \quad \textit{for all } u \in H_A.$$

In particular, $H_A \subset L^{\alpha(p)}(0,1)$.

Recall that in the case of profiles with tapering of order $p \le 2$, we have that $H_A \subset L^2(0,1)$.

Proof. For $2 < p < 3$, we use estimates (14) and (5). For all $u \in H_p$ and for all $s \in (0,1]$, we have

$$|u(s)| \le \|u\|_p \left\{ \frac{1 - s^{1-p}}{1-p} \right\}^{1/2} \le \frac{1}{\sqrt{K_2}} \|u\|_A \left\{ \frac{s^{1-p}}{p-1} \right\}^{1/2}.$$

For all $u \in H_A$, for all $\alpha > 1$, and if $(1-p)\frac{\alpha}{2} > -1$, we have

$$\begin{aligned} \int_0^1 |u(s)|^\alpha ds &\le \frac{\|u\|_A^\alpha}{(K_2(p-1))^{\alpha/2}} \int_0^1 s^{(1-p)\alpha/2} ds \\ &= \frac{\|u\|_A^\alpha}{(K_2(p-1))^{\alpha/2}} \cdot \left. \frac{s^{1+(1-p)\frac{\alpha}{2}}}{1 + (1-p)\frac{\alpha}{2}} \right|_0^1 \\ &= \tilde{C} \|u\|_A^\alpha < \infty. \end{aligned}$$

Now $1 + (1-p)\frac{\alpha}{2} > 0 \Longleftrightarrow \alpha < \frac{2}{p-1}$. But since $p < 3$, $\frac{2}{p-1} > 1$ and there exists $\alpha(p) \in (1, \frac{2}{p-1})$. With this $\alpha(p)$ we have

$$\|u\|_{L^{\alpha(p)}} = \left\{ \int_0^1 |u(s)|^{\alpha(p)} ds \right\}^{1/\alpha(p)} \le \tilde{C}^{1/\alpha(p)} \|u\|_A = C \|u\|_A.$$

Note that $p > 2$ implies that $\alpha(p) < \frac{2}{p-1} < 2$. □

Later on, we will need the following lemma, which is not difficult to check.

LEMMA 2.3. *For all $\theta \in \mathbb{R}$ and for all $\alpha \in [1, 2]$, we have*

$$|1 - \cos\theta| \le 2|\theta|^{\alpha},$$
$$|\theta - \sin\theta| \le 2|\theta|^{\alpha}.$$

2.2. Genus of a set. In this part, we recall a result due to Clark [2] concerning the existence of a finite or infinite number of critical points of a C^1-functional on a real Hilbert space $(H, \langle ., . \rangle)$. (See also [4] and [7].) It is based on the notion of the genus of a set. Let $M > 0$ and let

$$\Sigma = \big\{\Omega \subset H : \Omega \text{ is closed and } \Omega = -\Omega\big\},$$
$$\Sigma_M = \big\{\Omega \in \Sigma : \|u\| \le M \text{ for all } u \in \Omega\big\},$$

and define the genus $g : \Sigma \to \mathbb{N} \cup \{0, \infty\}$ as follows:

- $g(\emptyset) = 0$,
- $g(\Omega) = k$ if there is an odd mapping $h \in C(\Omega, \mathbb{R}^k \setminus \{0\})$ and k is the smallest integer with this property, and
- $g(\Omega) = \infty$ if there is no integer k with the above property.

Set

$$G_k = \Big\{\Omega \in \Sigma : g(\Omega) \ge k\Big\}.$$

We recall that if there is an odd homeomorphism from Ω onto the unit sphere on $\mathbb{R}^k$, we have $g(\Omega) = k$. Then we have that $G_k \neq \emptyset$ for all $k \in \mathbb{N}$ when $\dim H = \infty$. We will need the following result.

LEMMA 2.4. *Let $(H, \langle ., . \rangle)$ be a real Hilbert space and let $\Omega \in G_{i+1}$ for some $i \in \mathbb{N}^*$. If $P : H \to H_1$ is a continuous linear projection onto an i-dimensional subspace H_1 of H, then $\Omega \cap (I - P)(H) \neq \emptyset$.*

Proof. See Corollary 44.12 of [10]. □

Now we recall the definition of the condition of Palais and Smale (PS). A functional $f \in C^1(H, \mathbb{R})$ is said to satisfy the condition (PS) on H provided that every sequence $\{w_n\} \subset H$ which has the properties

(i) $\{f(w_n)\}$ is a bounded sequence,

(ii) $\|\nabla f(w_n)\| \to 0$

has a convergent subsequence in H.

We now define the following quantities. For $k \le \dim H$ and for $f \in C^1(H, \mathbb{R})$, let

$$b_k(f) = \inf_{\Omega \in G_k} \sup_{x \in \Omega} f(x). \tag{19}$$

Clearly $-\infty \le b_1(f) \le b_2(f) \le \cdots \le b_k(f) \le \cdots$.

The result of Clark is then the following.

THEOREM 2.5. *Let $f \in C^1(H, \mathbb{R})$ be an even functional with $f(0) = 0$, which is bounded below and satisfies the condition* (PS). *Suppose that* $\dim H = \infty$ *and that $-\infty < b_k(f) < 0$ for some $k \in \mathbb{N}^*$. Setting $K_b = \{w \in H : f(w) = b \text{ and } \nabla f(w) = 0\}$, we have that $K_{b_k(f)}$ is nonempty and compact. Moreover, if $-\infty < b_k(f) = b_{k+1}(f) = \cdots = b_{k+j-1}(f) < 0$, then $g(K_{b_k(f)}) \ge j$.*

In particular, if $-\infty < b_k(f) < 0$ for all $k \in \mathbb{N}^*$, we have that $K_{b_k(f)} \neq \emptyset$ for all $k \in \mathbb{N}^*$ and that f has an infinite number of critical points. Furthermore, $\lim_{k \to \infty} b_k(f) = 0$.

This theorem is due to Clark [2], except for the conclusion that $b_k(f) \to 0$ if $k \to \infty$, which is due to Heinz [4].

As was done by Stuart in [8], we are going to introduce a functional $j_\mu : H_A \to \mathbb{R}$. This functional will satisfy the hypotheses of Theorem 2.5. Applying this theorem to this functional, we obtain critical points which are solutions of problem P and satisfy the additional condition that $|u(s)| < \pi$ for all $s \in (0, 1]$. Stuart introduced this functional in the case of a profile with tapering of order $p = 2$, but it is possible to extend this to the case $2 < p < 3$. We do that in the next part.

3. A useful functional. In this part, we introduce a functional in order to apply Theorem 2.5, as was done by Stuart [8] to show that bifurcation from the solution $u \equiv 0$ occurs at every $\mu \geq \Lambda_e(A)$ in the case of profiles with tapering of order $p = 2$. Our goal is to extend this result to the case of profiles with tapering of order $2 < p < 3$ by showing that every $\mu \geq 0$ is a bifurcation value (recall that we are in the case $\Lambda(A) = 0$; see (10)). All the results of this part, which we prove in the case of profiles with tapering of order $2 < p < 3$, are also true in the case of order $p = 2$. For this order, the reader can find the proofs in [8, Lemmas 5.4 and 5.6 and Corollary 5.5], so we omit them in this case.

Let A be a profile with tapering of order $2 \leq p < 3$. Set

$$h(\theta) = \begin{cases} \sin\theta & \text{for } \theta \in [-\pi, \pi], \\ 0 & \text{for } \theta \notin [-\pi, \pi] \end{cases}$$

and let

$$H(\theta) = \int_0^\theta h(\sigma)d\sigma \quad \text{for all } \theta \in \mathbb{R}.$$

Clearly h is Lipschitz continuous on $\mathbb{R}$ with Lipschitz constant 1, and $H(\theta) = 1 - \cos\theta$ for $\theta \in [-\pi, \pi]$ and $H(\theta) = 2$ for $\theta \notin [-\pi, \pi]$, implying that $H \in C^1(\mathbb{R})$ is even. We define the functionals φ and $j_\mu(u) : H_A \to \mathbb{R}$ by

$$\varphi(u) = \int_0^1 H(u(s))ds \quad \text{and} \quad j_\mu(u) = \frac{1}{2}\|u\|_A^2 - \mu\varphi(u).$$

For u, $v \in H_A$,

$$\left| \int_0^1 v(s)h(u(s))ds \right| \leq \int_0^1 |v(s)|ds \leq C_1\|v\|_A,$$

where C_1 is the constant defined in (18), and so, by the Riesz representation theorem, there is a unique element $D_A(u) \in H_A$ such that

$$\langle D_A(u), v\rangle_A = \int_0^1 v(s)h(u(s))ds$$

for all $v \in H_A$.

LEMMA 3.1. *Let A be a profile with tapering of order $2 \leq p < 3$. The functional $\varphi : H_A \to \mathbb{R}$ has the following properties:*

(i) $0 \leq \varphi(u) = \varphi(-u) \leq 2$ *for all* $u \in H_A$.

(ii) $\varphi \in C^1(H_A)$ *and* $\nabla\varphi = D_A$.

(iii) $\varphi : H_A \to \mathbb{R}$ *is weakly sequentially continuous and* $D_A : H_A \to H_A$ *is completely continuous.*

Proof. (i) The proof of (i) is clear.
(ii) For all $\theta, \eta \in \mathbb{R}$, we have

$$\begin{aligned} H(\theta+\eta) - H(\theta) - h(\theta)\eta &= \int_0^1 \frac{d}{dt} H(\theta + t\eta)dt - h(\theta)\eta \\ &= \int_0^1 \{h(\theta + t\eta) - h(\theta)\}\eta dt. \end{aligned}$$

On one hand, since h is Lipschitz continuous with constant 1, we have

$$|h(\theta + t\eta) - h(\theta)| \le |\theta + t\eta - \theta| \le t|\eta|,$$

which implies

$$|H(\theta+\eta) - H(\theta) - h(\theta)\eta| \le \int_0^1 t\eta^2 dt = \frac{\eta^2}{2} \quad \text{for all } \theta, \eta \in \mathbb{R}.$$

On the other hand, we have

$$|h(\theta + t\eta) - h(\theta)| \le 2,$$

which implies

$$|H(\theta+\eta) - H(\theta) - h(\theta)\eta| \le 2\int_0^1 |\eta| dt = 2|\eta| \quad \text{for all } \theta, \eta \in \mathbb{R}.$$

Combining these two inequalities, we have, for all $\alpha \in [1, 2]$,

$$|H(\theta+\eta) - H(\theta) - h(\theta)\eta| \le 2|\eta|^\alpha \quad \text{for all } \theta, \eta \in \mathbb{R}.$$

Now for all $u, v \in H_A$, and using $\alpha(p)$ given by Lemma 2.2, we have

$$\begin{aligned} &|\varphi(u+v) - \varphi(u) - \langle D_A(u), v\rangle_A| \\ &= \left| \int_0^1 H(u(s) + v(s)) - H(u(s)) - h(u(s))v(s)ds \right| \\ &\le \int_0^1 |H(u(s) + v(s)) - H(u(s)) - h(u(s))v(s)|ds \\ &\le 2\int_0^1 |v(s)|^{\alpha(p)} ds \\ &\le 2C^{\alpha(p)} \|v\|_A^{\alpha(p)}, \end{aligned}$$

where C is the constant given by Lemma 2.2 and then

$$\lim_{\|v\|_A \to 0} \frac{|\varphi(u+v) - \varphi(u) - \langle D_A(u), v\rangle_A|}{\|v\|_A} \le \lim_{\|v\|_A \to 0} 2C^{\alpha(p)} \|v\|_A^{\alpha(p)-1} = 0,$$

showing that φ is Fréchet differentiable at u and $\varphi'(u)v = \langle D_A(u), v\rangle_A$ for all $u, v \in H_A$.

Now we shall show that $D_A : H_A \to H_A$ is completely continuous, which implies that φ' is continuous. Consider a sequence $\{u_n\}$ such that u_n converges weakly to

u in H_A. We must show that $\{D_A(u_n)\}$ converges strongly to $D_A(u)$ in H_A. For $v \in H_A$ and for $\varepsilon \in (0,1)$, we have

$$\begin{aligned}
|\langle D_A(u_n) - D_A(u), v\rangle_A| &= \left| \int_0^1 v(s)\{h(u_n(s)) - h(u(s))\}ds \right| \\
&\le 2\int_0^\varepsilon |v(s)|ds + \int_\varepsilon^1 |v(s)| \cdot |u_n(s) - u(s)|ds \\
&\le 2\|v\|_p \int_0^\varepsilon \left\{ \frac{s^{1-p}-1}{p-1} \right\}^{1/2} ds + \int_\varepsilon^1 |v(s)| \cdot |u_n(s) - u(s)|ds \\
&\le 2\frac{\|v\|_A}{\sqrt{K_2}} S(\varepsilon) + \int_\varepsilon^1 |v(s)| \cdot |u_n(s) - u(s)|ds,
\end{aligned}$$

where we used estimates (14) and (5) and we set

$$S(\varepsilon) = \int_0^\varepsilon \left\{ \frac{s^{1-p}-1}{p-1} \right\}^{1/2} ds.$$

Now since $u_n \to u$ uniformly on $[\varepsilon, 1]$, for $\eta > 0$, there exists $N(\eta) \in \mathbb{N}$ such that $n \ge N(\eta)$ implies $|u_n(s) - u(s)| < \eta$ for all $s \in [\varepsilon, 1]$. Thus, for $n \ge N(\eta)$, $|u_n - u|$ belongs to $L^\beta(0,1)$ for every $\beta \ge 1$. Choose β such that $1/\beta + 1/\alpha(p) = 1$, where $\alpha(p)$ is given by Lemma 2.2. We then have, for $n \ge N(\eta)$,

$$\begin{aligned}
&\int_\varepsilon^1 |v(s)| \cdot |u_n(s) - u(s)|ds \\
&\le \left\{ \int_\varepsilon^1 |v(s)|^{\alpha(p)} ds \right\}^{1/\alpha(p)} \left\{ \int_\varepsilon^1 |u_n(s) - u(s)|^\beta ds \right\}^{1/\beta} \\
&\le C\|v\|_A \cdot \eta,
\end{aligned}$$

where we have used the constant C given by Lemma 2.2. Thus we have

$$\limsup_{n\to\infty} \|D_A(u_n) - D_A(u)\|_A \le \frac{2}{\sqrt{K_2}} S(\varepsilon) \quad \text{for all } \varepsilon \in (0,1).$$

Furthermore, we have

$$0 \le \lim_{\varepsilon\to 0} S(\varepsilon) \le \frac{1}{\sqrt{p-1}} \lim_{\varepsilon\to 0} \int_0^\varepsilon s^{\frac{1-p}{2}} ds = 0$$

since we are in the case $2 < p < 3$. Then $\lim_{\varepsilon\to 0} S(\varepsilon) = 0$ and D_A is completely continuous.

(iii) We only have to prove that φ is weakly sequentially continuous. Consider a sequence $\{u_n\}$ which is weakly convergent to $u \in H_A$. For any $\varepsilon \in (0,1)$, we have

$$\begin{aligned}
|\varphi(u_n) - \varphi(u)| &= \left| \int_0^1 H(u_n(s)) - H(u(s))ds \right| \\
&\le 4\varepsilon + \int_\varepsilon^1 |H(u_n(s)) - H(u(s))|ds.
\end{aligned}$$

Since u_n converges uniformly to u on $[\varepsilon, 1]$, it follows that

$$\limsup_{n\to\infty} |\varphi(u_n) - \varphi(u)| \le 4\varepsilon \quad \text{for all } \varepsilon \in (0,1).$$

Then $\varphi(u_n)$ converges to $\varphi(u)$ and φ is weakly sequentially continuous. □

COROLLARY 3.2. *Let A be a profile with tapering of order $2 \le p < 3$. For all $\mu > 0$, the functional $j_\mu : H_A \to \mathbb{R}$ has the following properties:*

(i) *$j_\mu \in C^1(H_A)$ and $\nabla j_\mu = I - \mu D_A$.*

(ii) *j_μ is bounded below and satisfies the condition* (PS).

Proof. By Lemma 3.1, $j_\mu \in C^1(H_A)$ and $\nabla j_\mu(u) = I - \mu D_A$. Moreover, $j_\mu(u) \ge -2\mu$ for all $u \in H_A$. Consider a sequence $\{w_n\} \subset H_A$ such that

(i) $\{j_\mu(w_n)\}$ is bounded,

(ii) $\|\nabla j_\mu(w_n)\|_A \to 0$ as $n \to \infty$.

Since $j_\mu(u) = \frac{1}{2}\|u\|_A - \mu\varphi(u)$ and $0 \le \varphi(u) \le 2$ for all $u \in H_A$, it follows immediately from (i) that $\{w_n\}$ is a bounded sequence in H_A. Passing to a subsequence we can suppose that $w_n \rightharpoonup w$ weakly in H_A, and hence $\|D_A(w_n) - D_A(w)\|_A \to 0$ by Lemma 3.1(iii). But then

$$w_n = \nabla j_\mu(w_n) + \mu D_A(w_n) \to \mu D_A(w),$$

proving that the condition (PS) is satisfied. □

LEMMA 3.3. *Let A be a profile with tapering of order $2 \le p < 3$, and suppose that $\nabla j_\mu(u) = 0$ for some $\mu > 0$ and $u \in H_A$. Then u is a solution of problem P and $|u(s)| < \pi$ for all $s \in (0, 1]$.*

Proof. Suppose that $u \in H_A$ and $\nabla j_\mu(u) = 0$. Thus we have $u = \mu D_A(u)$, and then $\langle u, v\rangle_A = \mu\langle D_A(u), v\rangle_A$ for all $v \in H_A$. Then

$$\int_0^1 A(s)u'(s)v'(s)ds = \mu \int_0^1 v(s)h(u(s))ds \quad \text{for all } v \in H_A.$$

It follows that $A(s)u'(s)$ admits a generalized derivative on $(0, 1)$ and that

$$\{A(s)u'(s)\}' = -\mu h(u(s)) \quad \text{a.e. on } (0, 1).$$

However, since $u \in H_A$, we know that $u \in C((0, 1])$ and hence $Au' \in C^1((0, 1])$. From the properties of A, this implies that $u \in C^1((0, 1])$. Moreover, $u \in H_A$ implies $u(1) = 0$. Let $v \in C^1([0, 1])$ be such that $v(1) = 0$ and $v(s) = 1$ for all $s \le 1/2$. Clearly $v \in H_A \cap L^1(0, 1)$ and, for any $\varepsilon \in (0, 1/2)$,

$$\begin{aligned} A(\varepsilon)u'(\varepsilon) &= -\int_\varepsilon^1 A(s)u'(s)v'(s)ds - \int_\varepsilon^1 \{A(s)u'(s)\}'v(s)ds \\ &= -\int_0^1 A(s)u'(s)v'(s)ds + \mu\int_\varepsilon^1 v(s)h(u(s))ds \\ &= -\mu\int_0^\varepsilon h(u(s))ds \end{aligned}$$

since $v' \equiv 0$ on $(0, 1/2)$ and $\{A(s)u'(s)\}' = -\mu h(u(s))$ on $(0, 1)$. Then

$$|A(\varepsilon)u'(\varepsilon)| \le \mu\varepsilon \quad \text{for } \varepsilon \in (0, 1/2)$$

and, in particular,

$$\lim_{s\to 0} A(s)u'(s) = 0.$$

We have shown that $u \in C^1((0, 1])$, $Au' \in C^1((0, 1])$, and

$$\{A(s)u'(s)\}' + \mu h(u(s)) = 0 \quad \text{for all } s \in (0, 1] \tag{20}$$

with $\lim_{s\to 0} A(s)u'(s) = 0$ and $u(1) = 0$.

Let us now show that $|u(s)| < \pi$ for all $s \in (0,1]$. Then we will have that $h(u(s)) = \sin u(s)$ for all $s \in (0,1]$ and u will be a solution of problem P.

Suppose that there is a point $s_0 \in (0,1)$ such that $u(s_0) > \pi$ and let (a,b) be a maximal interval on which $u > \pi$. Then $h(u(s)) = 0$ on (a,b) and so there is a constant c such that $A(s)u'(s) = c$ on (a,b). Since $u(1) = 0$ we must have $c < 0$. Indeed, if $c \geq 0$ and since $A(s) \geq 0$ for all $s \in [0,1]$, we would have that u is increasing on (a,b). We would then have $b = 1$ and $u(1) \geq \pi$, which is in contradiction to $u(1) = 0$. But $c < 0$ implies that u is strictly decreasing on (a,b) and hence that $a = 0$.

But then $\lim_{s\to 0} A(s)u'(s) = c \neq 0$, which contradicts an earlier assertion. Hence $u \leq \pi$ on $(0,1]$. Now if there is a point $s \in (0,1)$ such that $u(s) = \pi$, $u'(s) = 0$ and, consequently, $u \equiv \pi$ on $(0,1]$ by the uniqueness of the Cauchy problem for (20). This is in contradiction to $u(1) = 0$, so we can conclude $u(s) < \pi$ on $(0,1]$. Replacing u by $-u$, we see that $|u(s)| < \pi$ for all $s \in (0,1]$. This concludes the proof. □

4. Bifurcation for profiles A with tapering of order $2 < p < 3$. In this section we consider profiles A with tapering of order $2 < p < 3$. We want to prove Theorem 1.4. To do that we use arguments similar to those used by Stuart [8] to show that for a profile with tapering of order $p = 2$, every $\mu \geq \Lambda_e(A)$ is a bifurcation value.

In order to apply Theorem 2.5 to our problem we still need to estimate quantities $b_k(j_\mu)$ for the functional j_μ introduced in section 3, and for this the following result is crucial.

LEMMA 4.1. *Let A be a profile with tapering of order $2 < p < 3$. Given any $k \in \mathbb{N}^*$ and any $\varepsilon > 0$, there is a subspace E of $H_A \cap L^\infty(0,1)$ such that* $\dim E = k$ *and*

$$\int_0^1 A(s)u'(s)^2 ds \leq \varepsilon \int_0^1 u(s)^2 ds$$

for all $u \in E$.

Proof. Let $k \in \mathbb{N}^*$ and $\varepsilon > 0$. First of all, we note that the space $H_2 \subset H_p$ for all $p \geq 2$. Now consider $0 < \delta < 1$ and set

$$A_\delta(s) = \begin{cases} A(s) & \text{if } \delta \leq s \leq 1, \\ \frac{A(\delta)s^2}{\delta^2} & \text{if } 0 \leq s < \delta. \end{cases}$$

It is clear that $A_\delta \in C([0,1])$ and there exists $L_\delta \in (0,\infty)$ such that

$$\lim_{s\to 0} \frac{A_\delta(s)}{s^2} = \frac{A(\delta)}{\delta^2} = L_\delta.$$

Then A_δ is a profile with tapering of order 2. By Proposition 1.6, there exists $T_\delta : H_2 \to H_2$ such that

$$\langle T_\delta u, v\rangle_{A_\delta} = \langle u, v\rangle_{L^2} \quad \text{for all } u, v \in H_2.$$

Furthermore, by Theorems 1.9 and 1.10, we have

$$\max \sigma_e(T_\delta) = \frac{4}{L_\delta} = \frac{1}{\Lambda_e(A_\delta)}.$$

Now, since

$$\lim_{\delta\to 0} L_\delta = \lim_{\delta\to 0} \frac{A(\delta)}{\delta^2} = \lim_{\delta\to 0} \frac{A(\delta)}{\delta^p} \cdot \frac{\delta^p}{\delta^2} = 0,$$

we have $\lim_{\delta\to 0}\max\sigma_e(T_\delta)=+\infty$ and $\lim_{\delta\to 0}\Lambda_e(A_\delta)=0$.

Then there exists $\delta_0>0$ such that $0<\delta<\delta_0$ implies

$$\Lambda_e(A_\delta)<\frac{K_2\varepsilon}{2K_1},$$

where the constants K_1 and K_2 are defined in (1). Choose $0<\delta<\delta_0$. Applying Lemma 5.7 of [8], we know that there is a subspace E of $H_2\cap L^\infty(0,1)$ such that dim $E=k$ and

$$\begin{aligned}\int_0^1 A_\delta(s)u'(s)^2ds &\le \Big\{\Lambda_e(A_\delta)+\frac{K_2\varepsilon}{2K_1}\Big\}\int_0^1 u(s)^2ds\\ &\le \Big\{\frac{K_2\varepsilon}{2K_1}+\frac{K_2\varepsilon}{2K_1}\Big\}\int_0^1 u(s)^2ds=\frac{K_2}{K_1}\varepsilon\int_0^1 u(s)^2ds\end{aligned}$$

for all $u\in E\subset H_2\cap L^\infty(0,1)$. Now since $H_2\subset H_p$, we have that $E\subset H_p\cap L^\infty(0,1)$. Then we only need to verify that

$$\int_0^1 A(s)u'(s)^2ds\le\varepsilon\int_0^1 u(s)^2ds\quad\text{for all } u\in E.$$

On one hand, for $0\le s<\delta$, we have

$$\begin{aligned}&A(s)\le K_1s^p\\ &=K_1\Big(\frac{s}{\delta}\Big)^{p-2}\cdot\frac{s^2}{\delta^2}\delta^p\le K_1\frac{s^2}{\delta^2}\delta^p\\ &\le\frac{K_1A(\delta)s^2}{K_2\delta^2}=\frac{K_1}{K_2}A_\delta(s).\end{aligned}$$

On the other hand, for $\delta\le s\le 1$, since $K_1/K_2\ge 1$ we have

$$A(s)=A_\delta(s)\le\frac{K_1}{K_2}A_\delta(s).$$

Finally we have shown that

$$A(s)\le\frac{K_1}{K_2}A_\delta(s)\quad\text{for all } s\in[0,1].$$

Then, for all $u\in E$, we have

$$\begin{aligned}&\int_0^1 A(s)u'(s)^2ds\le\frac{K_1}{K_2}\int_0^1 A_\delta(s)u'(s)^2ds\\ &\le\frac{K_1}{K_2}\cdot\frac{K_2}{K_1}\varepsilon\int_0^1 u(s)^2ds\\ &=\varepsilon\int_0^1 u(s)^2ds.\qquad\square\end{aligned}$$

LEMMA 4.2. *Let A be a profile with tapering of order $2<p<3$ and consider $\xi>0$. Let $v\in C^1((0,1])$ such that $Av'\in C^1((0,1])$. Suppose that v is any nontrivial solution of the linearized equation*

$$\big\{A(s)v'(s)\big\}'+\xi v(s)=0\quad\text{on } (0,1).\tag{21}$$

For any $n \in \mathbb{N}$, there exists $\delta > 0$ such that v has at least $n+1$ zeros in the interval $(\delta, 1]$.

Proof. Choose $\varepsilon > 0$ such that $\xi/\varepsilon > 1/4$. Since there exists $L \in (0, \infty)$ such that $\lim_{s\to 0} A(s)/s^p = L$, we have

$$\lim_{s\to 0} \frac{A(s)}{s^2} = \lim_{s\to 0} \frac{A(s)}{s^p} \cdot \frac{s^p}{s^2} = 0.$$

Then there exists $\eta > 0$ such that $s \in (0, \eta)$ implies $|A(s)/s^2| < \varepsilon$, that is, $A(s) < \varepsilon s^2$. Consider $w \neq 0$ a solution of the equation

$$\{\varepsilon s^2 w'(s)\}' + \xi w(s) = 0 \quad \text{on } (0,1).$$

Since $\xi/\varepsilon > 1/4$, as was shown by Stuart (see the proof of Corollary 5.2 of [9]), w has a sequence of zeros converging to 0. Now, by the Sturm comparison theorem (see Theorem 3.1 in Chapter II of [6]), every solution of (21) has at least one zero between successive zeros of w in $(0, \eta)$. Then there exists $0 < \delta < \eta$ such that v has at least $n+1$ zeros on (δ, η) and hence on $(\delta, 1]$. □

We now are able to prove Theorem 1.4.

Proof of Theorem 1.4. Consider $\mu > 0$. We apply Theorem 2.5 to the functional $j_\mu : H_A \to \mathbb{R}$. By Corollary 3.2 and Lemma 3.3, it is sufficient to show that $b_k(j_\mu) < 0$ for each $k \in \mathbb{N}^*$. To do that, consider $k \in \mathbb{N}^*$ and $\varepsilon > 0$ such that $\varepsilon < \mu$. Let E be the subspace given by Lemma 4.1 and, for $t > 0$, consider

$$\Omega_t = \Big\{ u \in E : \|u\|_{L^\infty} = t \Big\}.$$

Then the genus of Ω_t is equal to $k = \dim E$, and since $\|.\|_{L^\infty}$ and $\|.\|_A$ are equivalent on E, there exists $C > 0$ such that $\|u\|_A \geq Ct$ for all $u \in \Omega_t$. Fix $\delta \in (0, 1 - \varepsilon/\mu)$ and fix $t \in (0, \pi)$ such that

$$1 - \cos\theta \geq \frac{1-\delta}{2}\theta^2 \quad \text{for all } |\theta| \leq t.$$

Using this and Lemma 4.1, we have that for $u \in \Omega_t$,

$$\begin{aligned} j_\mu(u) &= \frac{1}{2}\|u\|_A^2 - \mu \int_0^1 \{1 - \cos u(s)\} ds \\ &\leq \frac{1}{2}\left\{ \|u\|_A^2 - \mu(1-\delta) \int_0^1 u(s)^2 ds \right\} \\ &\leq \frac{1}{2}\|u\|_A^2 \left\{ 1 - \mu(1-\delta)\frac{1}{\varepsilon} \right\} \\ &\leq \frac{(Ct)^2}{2}\left\{ 1 - \frac{(1-\delta)\mu}{\varepsilon} \right\} < 0 \end{aligned}$$

and thus

$$0 > \sup_{u\in\Omega_t} j_\mu(u) \geq \inf_{\Omega\in G_k} \sup_{u\in\Omega} j_\mu(u) = b_k(j_\mu).$$

The existence of a sequence $\{u_k\}$ of solutions of problem P with $j_\mu(u_k) = b_k(j_\mu)$ now follows from Theorem 2.5 and Lemma 3.3. Furthermore, since $\lim_{k\to\infty} b_k(j_\mu) = 0$,

we have $j_\mu(u_k) \to 0$ and $\nabla j_\mu(u_k) = 0$, so the condition (PS) implies that $\{u_k\}$ has a subsequence $\{u_{k_i}\}$ which converges to an element u in H_A. Then $j_\mu(u) = 0$, $\nabla j_\mu(u) = 0$, and $|u(s)| \le \pi$ for all $s \in (0,1]$ since $\{u_{k_i}\}$ converges to u uniformly on compact subsets of $(0,1]$. By Lemma 3.3 we can conclude that $|u(s)| < \pi$ for all $s \in (0,1]$. However,

$$\begin{aligned} 0 &= 2j_\mu(u) - \langle \nabla j_\mu(u), u\rangle_A \\ &= \|u\|_A^2 - \mu \int_0^1 \{2 - 2\cos u(s)\} ds - \{\langle u - \mu D_A(u), u\rangle_A\} \\ &= \mu \int_0^1 \{u(s)\sin u(s) - 2 + 2\cos u(s)\} ds \end{aligned}$$

and $\theta \sin\theta - 2\{1 - \cos\theta\} < 0$ for $0 < |\theta| < \pi$. This implies that $u \equiv 0$ on $(0,1]$. Since this argument applies to every subsequence of $\{u_k\}$, we can conclude that the whole sequence $\{u_k\}$ converges to 0 in H_A. Indeed, if $\{u_k\}$ does not converge in H_A, there exists a constant $\eta > 0$ and a subsequence $\{u_{k_i}\}$ such that $\|u_{k_i}\|_A \ge \eta$ for all i. But this subsequence satisfies $j_\mu(u_{k_i}) \to 0$ if $i \to \infty$ and $\nabla j_\mu(u_{k_i}) = 0$ for all i. Thus, by the condition (PS) and repeating the argument used above, $\{u_{k_i}\}$ has a convergent subsequence to 0 in H_A. Now this is a contradiction.

Now fix $n \in \mathbb{N}$. We show that there exists $K \in \mathbb{N}$ such that u_k has at least n zeros in $(0,1]$ for all $k \ge K$. First choose $\xi \in (0,\mu)$ and any nontrivial solution v of the linearized equation

$$\big\{A(s)v'(s)\big\}' + \xi v(s) = 0 \quad \text{on } (0,1).$$

By Lemma 4.2, there exists $\delta > 0$ such that v has at least $n+1$ zeros in the interval $(\delta, 1]$. Since the sequence $\{u_k\}$ tends to 0 in H_A, it converges to 0 uniformly on $[\delta, 1]$, and so there is a constant $K \in \mathbb{N}$ such that

$$q_k(s) = \mu \frac{\sin u_k(s)}{u_k(s)} > \xi \quad \text{for all } s \in [\delta,1] \text{ and all } k \ge K.$$

But $\{u_k\}$ satisfies the linear equation

$$\big\{A(s)v'(s)\big\}' + q_k(s)v(s) = 0 \quad \text{on } (0,1)$$

and so, by the Sturm comparison theorem (see Theorem 3.1 in Chapter II of [6]), u_k vanishes at least once between successive zeros of v in $(\delta,1]$. Hence u_k has at least n zeros in $(\delta,1]$ for all $k \ge K$. □

5. Profile of order $p = 2$: Bifurcation at simple eigenvalues. The goal of this part is to prove Theorems 1.11 and 1.12. We use arguments similar to those of the previous section. We first need some lemmas.

LEMMA 5.1. *Let $(X, \|.\|)$ be a normed space. If $\{v_1, \dots, v_k\}$ are linearly independent, there is an $\varepsilon > 0$ with the following property: If $\{w_1, \dots, w_k\} \subset X$ is such that $\|v_i - w_i\| \le \varepsilon$ for all $i = 1, \dots, k$, then $\{w_1, \dots, w_k\}$ is linearly independent.*

Proof. Suppose that there is no such $\varepsilon > 0$. Then, for each $n \in \mathbb{N}^*$, there exists a set $\{u_1^n, \dots, u_k^n\}$ which is linearly dependent and such that $\|u_i^n - v_i\| \le 1/n$ for all $i = 1, \dots, k$. Then we can find $\lambda_1^n, \dots, \lambda_k^n$ such that

$$\sum_{i=1}^k \lambda_i^n u_i^n = 0$$

and

$$\lambda^n = \left\{ \sum_{i=1}^{k} (\lambda_i^n)^2 \right\}^{1/2} > 0.$$

Then we have

$$\sum_{i=1}^{k} \frac{\lambda_i^n}{\lambda^n} u_i^n = 0.$$

Now, for $i = 1, \dots, k$ and since $|\lambda_i^n|/\lambda^n \leq 1$ for all $n \in \mathbb{N}^*$, passing to a subsequence, there exists γ_i such that $\lambda_i^n/\lambda^n \to \gamma_i$ as $n \to \infty$. We then have

$$\sum_{i=1}^{k} \gamma_i v_i = 0.$$

Since $\{v_1, \dots, v_k\}$ is linearly independent, we have $\gamma_i = 0$ for all $i = 1, \dots, k$. But, since for all $n \in \mathbb{N}^*$,

$$\sum_{i=1}^{k} \left\{ \frac{\lambda_i^n}{\lambda^n} \right\}^2 = \frac{\sum_{i=1}^{k} (\lambda_i^n)^2}{(\lambda^n)^2} = 1,$$

and since

$$\lim_{n \to \infty} \sum_{i=1}^{k} \left\{ \frac{\lambda_i^n}{\lambda^n} \right\}^2 = \sum_{i=1}^{k} \gamma_i^2,$$

we have $\sum_{i=1}^{k} \gamma_i^2 = 1$. Now this is a contradiction. □

LEMMA 5.2. *Let H be a Hilbert space and let $S : D(S) \to H$, with $D(S) \subset H$, be self-adjoint and bounded below. We suppose that all eigenvalues of S are simple. We set $\lambda_e := \inf\{\lambda : \lambda \in \sigma_e(S)\}$, where $\sigma_e(S)$ denotes the essential spectrum of S, and let $\lambda_1 < \lambda_2 < \cdots < \lambda_k < \lambda_e$ be eigenvalues of S arranged in increasing order with corresponding orthonormal eigenvectors $e_1, e_2, \dots, e_k$. Then we have*

$$\begin{aligned} \inf \big\{ (S\psi, \psi) : \|\psi\| = 1, \psi \in D(S) \cap [e_1, \dots, e_k]^\perp \big\} \\ = \inf \big\{ \lambda : \lambda \in \sigma(S) \setminus \{\lambda_1, \dots, \lambda_k\} \big\}, \end{aligned}$$

where $\sigma(S)$ denotes the spectrum of S.

Proof. See Lemma 1.1 in Chapter XI of [3]. □

As a consequence of this lemma, we have the following.

LEMMA 5.3. *Under assumption* (H)*, and with the corresponding notation, choose $i \in I$. Then we have*

$$\|\psi\|_A^2 - \mu_i^+ \|\psi\|_{L^2}^2 \geq 0 \quad \textit{for all } \psi \in [\varphi_1, \dots, \varphi_i]^\perp.$$

Proof. We apply Lemma 5.2 to the operator $S = -T : H_A \to H_A$. Since T is self-adjoint, it is clear that S is self-adjoint. Moreover, using Lemma 2.1 and (5),

$$\langle Tu, u \rangle_A = \|u\|_{L^2}^2 \leq \frac{4}{K_2} \|u\|_A^2 \quad \text{for all } u \in H_A.$$

Then

$$\langle Su,u\rangle_A = -\langle Tu,u\rangle_A \geq -\frac{4}{K_2}\|u\|_A^2 \quad \text{for all } u \in H_A,$$

implying that S is bounded below. The spectrum of S is given by $\sigma(S) = -\sigma(T)$. Applying Lemma 5.2, we have

$$\begin{aligned}&\inf\big\{(S\psi,\psi) : \|\psi\| = 1, \psi \in H_A \cap [\varphi_1,\dots,\varphi_i]^\perp\big\}\\ &= \inf\big\{\lambda : \lambda \in \sigma(S) \setminus \{-\mu_1^{-1},\dots,-\mu_i^{-1}\}\big\} = -\frac{1}{\mu_i^+}.\end{aligned}$$

We then have

$$\begin{aligned}&\left\langle -T\Big(\frac{\psi}{\|\psi\|_A}\Big), \frac{\psi}{\|\psi\|_A}\right\rangle_A \geq -\frac{1}{\mu_i^+} \quad \text{for all } \psi \in [\varphi_1,\dots,\varphi_i]^\perp \setminus \{0\}\\ \Rightarrow\ &-\langle T\psi,\psi\rangle_A \geq -\frac{1}{\mu_i^+}\|\psi\|_A^2 \quad \text{for all } \psi \in [\varphi_1,\dots,\varphi_i]^\perp,\end{aligned}$$

which implies

$$\|\psi\|_A^2 - \mu_i^+\|\psi\|_{L^2}^2 \geq 0 \quad \text{for all } \psi \in [\varphi_1,\dots,\varphi_i]^\perp. \qquad \square$$

LEMMA 5.4. *Under assumption* (H)*, and with the corresponding notation, choose* $i \in I$*. Given any* $\varepsilon > 0$*, there is a subspace* E *of* $H_A \cap L^\infty(0,1)$ *such that* $\dim E = i$ *and*

$$\int_0^1 A(s)u'(s)^2 ds \leq (\mu_i + \varepsilon)\int_0^1 u(s)^2 ds$$

for all $u \in E$.

Proof. Let $\varepsilon > 0$. Set $F = span\{\varphi_1,\dots,\varphi_i\}$. Then for $j = 1,\dots,i$, we have

$$\begin{aligned}\int_0^1 A(s)\varphi_j'(s)^2 ds &= \langle \varphi_j,\varphi_j\rangle_A = \mu_j\langle T\varphi_j,\varphi_j\rangle_A\\ &= \mu_j\int_0^1 \varphi_j(s)^2 ds \leq \mu_i\int_0^1 \varphi_j(s)^2 ds.\end{aligned}$$

Then, for all $u = \sum_{j=1}^i \alpha_j\varphi_j \in F$, we have

$$\begin{aligned}&\int_0^1 A(s)\left(\sum_{j=1}^i \alpha_j\varphi_j\right)'(s)^2 ds = \left\langle \sum_{j=1}^i \alpha_j\varphi_j, \sum_{j=1}^i \alpha_j\varphi_j\right\rangle_A\\ &= \sum_{j=1}^i \langle \alpha_j\varphi_j,\alpha_j\varphi_j\rangle_A = \sum_{j=1}^i \mu_j\langle T\alpha_j\varphi_j,\alpha_j\varphi_j\rangle_A\\ &\leq \mu_i\sum_{j=1}^i \langle T\alpha_j\varphi_j,\alpha_j\varphi_j\rangle_A = \mu_i\left\langle T\left(\sum_{j=1}^i \alpha_j\varphi_j\right), \sum_{j=1}^i \alpha_j\varphi_j\right\rangle_A\\ &= \mu_i\int_0^1 u(s)^2 ds.\end{aligned}$$

We then have

$$\int_0^1 A(s)u'(s)^2 ds \le \mu_i \int_0^1 u(s)^2 ds \quad \text{for all } u \in F.$$

Now, since $H_A \cap L^\infty(0,1)$ is dense in H_A and T is continuous, there exist $v_1, \dots, v_i \in H_A \cap L^\infty(0,1)$ (which are linearly independent by Lemma 5.1) such that if $E = span\{v_1, \dots, v_i\}$, then $E \subset H_A \cap L^\infty(0,1)$, dim $E = i$, and

$$\int_0^1 A(s)v'(s)^2 ds \le (\mu_i + \varepsilon) \int_0^1 v(s)^2 ds \quad \text{for all } v \in E. \qquad \square$$

Now we are able to give the proof of Theorem 1.11.

Proof of Theorem 1.11. We apply Theorem 2.5 to the functional $j_\mu : H_A \to \mathbb{R}$. In light of Corollary 3.2 and Lemma 3.3, we only need to show that $b_i(j_\mu) < 0$ (recall that $b_k(j_\mu) \le b_i(j_\mu)$ for $k = 1, \dots, i$). To this end choose $\varepsilon > 0$ such that $\mu_i + \varepsilon < \mu$. Let E be the subspace given by Lemma 5.4 and, for $t > 0$, let

$$\Omega_t = \Big\{ u \in E : \|u\|_{L^\infty} = t \Big\}.$$

Then the genus of Ω_t is equal to $i = \dim E$, and since $\|.\|_{L^\infty}$ and $\|.\|_A$ are equivalent on E, there exists $C > 0$ such that $\|u\|_A \ge Ct$ for all $u \in \Omega_t$. Fix $\delta \in (0, 1 - (\mu_i + \varepsilon)/\mu)$ and fix $t \in (0, \pi)$ such that

$$1 - \cos\theta \ge \frac{1-\delta}{2}\theta^2 \quad \text{for all } |\theta| \le t.$$

Using this and Lemma 5.4, we have that for $u \in \Omega_t$,

$$\begin{aligned} j_\mu(u) &= \frac{1}{2}\|u\|_A^2 - \mu \int_0^1 \{1 - \cos u(s)\} ds \\ &\le \frac{1}{2}\Big\{ \|u\|_A^2 - \mu(1-\delta) \int_0^1 u(s)^2 ds \Big\} \\ &\le \frac{1}{2}\|u\|_A^2 \Big\{ 1 - \mu(1-\delta)\frac{1}{\mu_i + \varepsilon} \Big\} \\ &\le \frac{(Ct)^2}{2} \Big\{ 1 - \frac{(1-\delta)\mu}{\mu_i + \varepsilon} \Big\} < 0 \end{aligned}$$

and thus

$$0 > \sup_{u \in \Omega_t} j_\mu(u) \ge \inf_{\Omega \in G_i} \sup_{u \in \Omega} j_\mu(u) = b_i(j_\mu).$$

The existence of a set $\{u_k\}$, $k = 1, \dots, i$, of solutions of problem P with $j_\mu(u_k) = b_i(j_\mu)$ now follows from Theorem 2.5 and Lemma 3.3. $\square$

Using the notation of section 2.2, we have the following.

LEMMA 5.5. *Under assumption* (H)*, and with the corresponding notation, choose* $i \in I$. *Let* $\mu_i < \mu < \mu_i^+$ *and consider* $b_i(j_\mu)$. *Then there exists* $M = \sqrt{2 + 4\Lambda_e(A)} > 0$ *(independent of* μ*) with the following property: For each* $0 < \varepsilon \le 1$, *there exists* $\Omega_\varepsilon \in \Sigma_M$ *such that*

(i) $g(\Omega_\varepsilon) \ge i$,

(ii) $b_i(j_\mu) \le \sup_{u \in \Omega_\varepsilon} j_\mu(u) < b_i(j_\mu) + \varepsilon$.

Remark. Lemma 5.5 means that to construct $b_i(j_\mu)$ for $\mu \in (\mu_i, \mu_i^+)$, it is sufficient to consider the sets $\Omega \in G_i$ such that $\Omega \in \Sigma_M$, instead of all sets $\Omega \in G_i$ such that $\Omega \in \Sigma$.

Proof. Let $0 < \varepsilon \leq 1$. By definition of the infimum, there is $\Omega \in G_i$ such that

$$b_i(j_\mu) \leq \sup_{u\in\Omega} j_\mu(u) < b_i(j_\mu) + \varepsilon.$$

Then, for each $u \in \Omega$,

$$\frac{1}{2}\|u\|_A^2 - \mu\varphi(u) < b_i(j_\mu) + \varepsilon,$$

which implies, since $0 \leq \varphi(u) \leq 2$ for all $u \in H_A$,

$$\|u\|_A^2 < 2(b_i(j_\mu) + 1) + 4\mu < 2 + 4\Lambda_e(A),$$

where we used the fact proved by Theorem 1.11 that $b_i(j_\mu) < 0$. Setting $M = \sqrt{2 + 4\Lambda_e(A)}$, we have that $\Omega \in \Sigma_M$. □

LEMMA 5.6. *Under assumption* (H)*, and with the corresponding notation, choose* $i \in I$. *Let* $\mu_i < \mu < \mu_i^+$. *Then we have* $b_{i+1}(j_\mu) \geq 0$.

Proof. For each $\varepsilon > 0$, there is $\Omega \in G_{i+1}$ such that

$$b_{i+1}(j_\mu) \leq \sup_{u\in\Omega} j_\mu(u) < b_{i+1}(j_\mu) + \varepsilon.$$

By Lemma 2.4, there exists $\psi \in \Omega \cap [\varphi_1, \ldots, \varphi_i]^\perp$. By Lemma 5.3, we have

$$\|\psi\|_A^2 - \mu_i^+\|\psi\|_{L^2}^2 \geq 0.$$

We then have

$$\begin{aligned} j_\mu(\psi) &= \frac{1}{2}\|\psi\|_A^2 - \mu\varphi(\psi) \\ &\geq \frac{1}{2}\|\psi\|_A^2 - \mu\frac{1}{2}\int_0^1 \psi(s)^2 ds \\ &\geq \frac{1}{2}\{\|\psi\|_A^2 - \mu_i^+\|\psi\|_{L^2}^2\} \geq 0. \end{aligned}$$

Then $\sup_{u\in\Omega} j_\mu(u) \geq 0$ and $b_{i+1}(j_\mu) \geq 0$. □

PROPOSITION 5.7. *Under assumption* (H)*, and with the corresponding notation, choose* $i \in I$. *We have*

$$\lim_{\mu\to\mu_{i+}} b_i(j_\mu) = 0.$$

Proof. By definition, and using Lemma 5.5, we have

$$b_i(j_\mu) = \inf_{\Omega\in G_i}\sup_{u\in\Omega} j_\mu(u) = \inf_{\Omega\in G_i\cap\Sigma_M}\sup_{u\in\Omega} j_\mu(u).$$

Let $\Omega \in G_i \cap \Sigma_M$, and let $u \in \Omega$. Then

$$\begin{aligned} j_\mu(u) &= \frac{1}{2}\|u\|_A^2 - \mu\varphi(u) \\ &\geq \frac{1}{2}\{\|u\|_A^2 - \mu\|u\|_{L^2}^2\} \\ &= \frac{1}{2}\{\|u\|_A^2 - \mu_i\|u\|_{L^2}^2 + (\mu_i - \mu)\|u\|_{L^2}^2\}. \end{aligned}$$

Since $g(\Omega) \geq i$, by Lemma 2.4, there exists $\psi \in \Omega \cap [\varphi_1, \ldots, \varphi_{i-1}]^{\perp}$. By Lemma 5.3, we have

$$\|\psi\|_A^2 - \mu_i \|\psi\|_{L^2}^2 \geq 0.$$

Using Lemma 2.1 and (5), we then have (recalling that $\psi \in \Sigma_M$),

$$j_\mu(\psi) \geq \frac{1}{2}(\mu_i - \mu)\|\psi\|_{L^2}^2 \geq \frac{2}{K_2} M^2(\mu_i - \mu).$$

Thus we have shown that for each $\Omega \in G_i \cap \Sigma_M$, there is a $\psi \in \Omega$ such that

$$j_\mu(\psi) \geq \frac{2}{K_2} M^2(\mu_i - \mu).$$

Then, for all $\mu \in (\mu_i, \mu_i^+)$, we have

$$0 > b_i(j_\mu) \geq \frac{2}{K_2} M^2(\mu_i - \mu) \to 0 \quad \text{if } \mu \to \mu_{i+}.$$

Then $\lim_{\mu \to \mu_{i+}} b_i(j_\mu) = 0$. □

LEMMA 5.8. *Let there be a sequence $\{\mu_k\} \subset \mathbb{R}_+$ which converges to $\mu \in \mathbb{R}$. Consider a sequence $\{u_k\} \subset H_A$ such that*

(i) *$\{j_{\mu_k}(u_k)\}$ is bounded;*

(ii) *$\|\nabla j_{\mu_k}(u_k)\|_A \to 0$ if $k \to \infty$.*

Then $\{u_k\}$ has a convergent subsequence.

Proof. Since $j_\mu(u) = \frac{1}{2}\|u\|_A^2 - \mu\varphi(u)$ with $0 \leq \varphi(u) \leq 2$, and since $\{j_{\mu_k}(\mu_k)\}$ is bounded, we have that $\{u_k\}$ is a bounded sequence in H_A. Then there is a subsequence $\{u_{k_i}\}$, and $u \in H_A$ such that u_{k_i} converges weakly to u in H_A. But since D_A is completely continuous, we have $\|D_A(u_{k_i}) - D_A(u)\|_A \to 0$ if $i \to \infty$.

Now, since $\nabla j_\mu = I - \mu D_A$, we have

$$u_{k_i} = \nabla j_{\mu_{k_i}}(u_{k_i}) + \mu_{k_i} D_A(u_{k_i}) \to 0 + \mu D_A(u) \quad \text{if } i \to \infty.$$

Finally $\{u_{k_i}\}$ converges in H_A. □

We now prove our second bifurcation result.

Proof of Theorem 1.12. For $\mu > \mu_i$, by Theorem 1.11 there is $u_\mu \in H_A \setminus \{0\}$ such that u_μ is a solution of problem P, $|u_\mu(s)| < \pi$ for all $s \in (0, 1]$, $\nabla j_\mu(u_\mu) = 0$, and $j_\mu(u_\mu) = b_i(j_\mu)$. Now choose a sequence $\{\alpha_k\}$ such that $\alpha_k > \mu_i$ for all $k \geq 1$ and such that $\alpha_k \to \mu_i$ as $k \to \infty$. Then, for each $k \geq 1$, there exists $u_k \in H_A \setminus \{0\}$ such that $j_{\alpha_k}(u_k) = b_i(j_{\alpha_k})$ and u_k is a solution of problem P. Furthermore, $\nabla j_{\alpha_k}(u_k) = 0$ for all $k \geq 1$. Now Proposition 5.7 implies

$$\lim_{k \to \infty} j_{\alpha_k}(u_k) = \lim_{k \to \infty} b_i(j_{\alpha_k}) = 0.$$

Thus $\{j_{\alpha_k}(u_k)\}$ is bounded. By Lemma 5.8, there is a subsequence $\{u_{k_i}\}$ which converges to $u \in H_A$. We then have $j_{\mu_i}(u) = 0$, $\nabla j_{\mu_i}(u) = 0$, and $|u(s)| \leq \pi$ for all $s \in (0, 1]$ since $\{u_{k_i}\}$ converges to u uniformly on compact subsets of $(0, 1]$. By Lemma 3.3, we have $|u(s)| < \pi$ for all $s \in (0, 1]$. But we have

$$\begin{aligned}
0 &= 2j_{\mu_i}(u) - \langle \nabla j_{\mu_i}(u), u \rangle_A \\
&= 2j_{\mu_i}(u) - \langle u - \mu_i D_A(u), u \rangle_A \\
&= \|u\|_A^2 - \mu_i \int_0^1 \{2 - 2\cos u(s)\} ds - \langle u, u \rangle_A + \mu_i \langle D_A(u), u \rangle_A \\
&= \mu_i \int_0^1 \{u(s) \sin u(s) - 2 + 2\cos u(s)\} ds
\end{aligned}$$

and $\theta \sin\theta - 2\{1 - \cos\theta\} < 0$ for all $0 < |\theta| < \pi$. This implies that $u \equiv 0$ on $(0, 1]$. Since this argument applies to every subsequence of $\{u_k\}$, we can conclude that the whole sequence $\{u_k\}$ converges to 0 in H_A. □

Acknowledgment. I would like to thank Professor C. A. Stuart for his precious advice that helped me to do this work.

REFERENCES

[1] H. Brézis, *Analyse Fonctionnelle: Théorie et Applications*, Dunod, Paris, 1999.
[2] D. C. Clark, *A variant of the Ljusternik-Schnirelman theory*, Indiana Univ. Math. J., 22 (1972), pp. 65–74.
[3] D. E. Edmunds and W. D. Evans, *Spectral Theory and Differential Operators*, Oxford University Press, Oxford, 1987.
[4] H.-P. Heinz, *Free Ljusternik-Schnirelman theory and the bifurcation diagrams of certain singular nonlinear problems*, J. Differential Equations, 66 (1987), pp. 263–300.
[5] M. A. Krasnoselskii, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, London, 1964.
[6] W. T. Reid, *Sturmian Theory for Ordinary Differential Equations*, Springer-Verlag, Berlin, 1980.
[7] M. Struwe, *Variational Methods*, Springer-Verlag, Berlin, 1990.
[8] C. A. Stuart, *Buckling of a heavy tapered rod*, J. Math. Pures Appl., 80 (2001), pp. 281–337.
[9] C. A. Stuart, *On the spectral theory of a heavy tapered rod*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 729–764.
[10] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*. III. *Variational Methods and Optimization*, Springer-Verlag, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 34, No. 5, pp. 1152–1166

HIGH-ORDER TERMS IN THE ASYMPTOTIC EXPANSIONS OF THE STEADY-STATE VOLTAGE POTENTIALS IN THE PRESENCE OF CONDUCTIVITY INHOMOGENEITIES OF SMALL DIAMETER*

HABIB AMMARI† AND HYEONBAE KANG‡

Abstract. We derive high-order terms in the asymptotic expansions of the steady-state voltage potentials in the presence of a finite number of diametrically small inhomogeneities with conductivities different from the background conductivity. Our derivation is rigorous and based on layer potential techniques. The asymptotic expansions in this paper are valid for inhomogeneities with Lipschitz boundaries and those with extreme conductivities.

Key words. small conductivity inhomogeneities, asymptotic expansions, generalized polarization tensors

AMS subject classification. 35B30

PII. S0036141001399234

1. Introduction.

Let Ω be a bounded domain in $\mathbb{R}^d$, $d \geq 2$, with a connected Lipschitz boundary $\partial\Omega$. Let ν denote the unit outward normal to $\partial\Omega$. Suppose that Ω contains a finite number m of small inhomogeneities $(D_l)_{l=1}^m$, each of the form $D_l = z_l + \epsilon B_l$, where B_l, $l = 1, \ldots, m$, is a bounded Lipschitz domain in $\mathbb{R}^d$ containing the origin. We assume that the domains $(D_l)_{l=1}^m$ are separated from each other and from the boundary. More precisely, we assume that there exists a constant $c_0 > 0$ such that

$$|z_l - z_{l'}| \geq 2c_0 > 0 \quad \forall\, l \neq l' \quad \text{and} \quad \operatorname{dist}(z_l, \partial\Omega) \geq 2c_0 > 0 \quad \forall\, l, \tag{1.1}$$

that ϵ, the common order of magnitude of the diameters of the inhomogeneities, is sufficiently small, that these inhomogeneities are disjoint, and that their distance to $\mathbb{R}^d \setminus \overline{\Omega}$ is larger than c_0. We also assume that the "background" is homogeneous with conductivity 1 and the inhomogeneities D_l have conductivities k_l, $k_l \neq 1$, $1 \leq l \leq m$.

Let u_ϵ denote the steady-state voltage potential in the presence of the conductivity inhomogeneities, i.e., the solution to

$$\begin{cases} \nabla \cdot \left(\chi\left(\Omega \setminus \bigcup_{l=1}^m \overline{D_l} \right) + \sum_{l=1}^m k_l \chi(D_l) \right) \nabla u_\epsilon = 0 \quad \text{in } \Omega, \\ \dfrac{\partial u_\epsilon}{\partial \nu}\Big|_{\partial\Omega} = g. \end{cases} \tag{1.2}$$

*Received by the editors December 5, 2001; accepted for publication (in revised form) November 4, 2002; published electronically April 15, 2003. This paper was completed while both authors were at the Mathematical Sciences Research Institute (MSRI) during the special program on inverse problems.

http://www.siam.org/journals/sima/34-5/39923.html

†Centre de Mathématiques Appliquées, CNRS UMR 7641 & Ecole Polytechnique, 91128 Palaiseau Cedex, France (ammari@cmapx.polytechnique.fr). This author was partially supported by ACI Jeunes Chercheurs (0693) from the Ministry of Education and Scientific Research, France.

‡School of Mathematical Sciences, Seoul National University, Seoul 151-747, Korea (hkang@math.snu.ac.kr). This author was partially supported by KOSEF 98-0701-03-5 and BK21 at the School of Mathematical Sciences of SNU.

Let U denote the "background" potential, that is, the solution to

$$\begin{cases} \Delta U = 0 \quad \text{in } \Omega, \\ \dfrac{\partial U}{\partial \nu}\big|_{\partial\Omega} = g. \end{cases} \tag{1.3}$$

The function g represents the applied boundary current; it belongs to $L^2_0(\partial\Omega) = \{g \in L^2(\partial\Omega), \int_{\partial\Omega} g = 0\}$. The potentials, u_ϵ and U, are normalized by $\int_{\partial\Omega} u_\epsilon = \int_{\partial\Omega} U = 0$.

The main achievement of this paper is a rigorous derivation, based on layer potential techniques, of high-order terms in the asymptotic expansion of $u_\epsilon|_{\partial\Omega}$ as $\epsilon \to 0$. The leading order term in this asymptotic formula has been derived by Cedio-Fengya, Moskow, and Vogelius [7]; see also the prior work of Friedman and Vogelius [14] for the case of perfectly conducting or insulating inhomogeneities. The main result of this paper is the following full asymptotic expansion of the solution for the case $m = 1$.

THEOREM 1.1. *Suppose that the inhomogeneity consists of a single component and let u_ε be the solution of* (1.2). *The following pointwise asymptotic expansion on $\partial\Omega$ holds for $d = 2, 3$:*

$$\begin{aligned} u_\varepsilon(x) = U(x) &- \epsilon^{d-2} \sum_{|i|=1}^{n} \sum_{|j|=1}^{n-|i|+1} \frac{1}{j!} \epsilon^{|i|+|j|} \\ &\times \left[\left(\left(I + \sum_{p=1}^{n+2-|i|-|j|-d} \epsilon^{d+p-1} \mathcal{Q}_p \right) (\partial^l U(z)) \right)_i M_{ij} \partial_z^j N(x,z) \right] \\ &+ O(\epsilon^{d+n}), \end{aligned} \tag{1.4}$$

where the remainder $O(\epsilon^{d+n})$ is dominated by $C\epsilon^{d+n}\|g\|_{L^2(\partial\Omega)}$ for some C independent of $x \in \partial\Omega$. Here $N(x,z)$ is the Neumann function, that is, the solution to (2.12)–(2.13), *M_{ij}, $i, j \in \mathbb{N}^d$, are the generalized polarization tensors defined in* (3.2), *and the matrix $\mathcal{Q}_p$ is defined in* (4.12).

We have a similar expansion for the solutions of the Dirichlet problem (Theorem 4.2).

The derivation of the asymptotic expansions for any fixed number m of well-separated inhomogeneities (these are a fixed distance apart) follows by iteration of the arguments that we will present for the case $m = 1$. In other words, we may develop asymptotic formulas involving the difference between the fields u_ϵ and U on $\partial\Omega$ with l inhomogeneities and those with $l - 1$ inhomogeneities, $l = m, \ldots, 1$, and then at the end essentially form the sum of these m formulas (the reference fields change, but that may easily be remedied). The derivation of each of the m formulas is virtually identical.

We also note that the asymptotic expansion (1.4) is valid for inhomogeneities with zero or infinity conductivity (cavity or perfect conductor). Precise definitions of generalized polarization tensors (GPTs) associated with the domains B_l and the conductivities k_l will be given at the end of section 3. These GPTs seem to be natural generalizations of the tensors that have been introduced by Schiffer and Szegö [23] and thoroughly studied by many other authors [22], [18], [14], [7]. (See section 3.)

The higher-order terms are essential when $\nabla U(z_l) = 0$, for then the leading order term in the asymptotic expansion of $u_\epsilon|_{\partial\Omega}$, given in [7], vanishes. We remind the reader that, for general current inputs g, ∇U vanishes at some "critical points" inside Ω.

The proof of our asymptotic expansion is radically different from the ones in [14], [7], and [26]. It is based on layer potential techniques and a decomposition formula of the steady-state voltage potential into a harmonic part and a refraction part. This formula is due to Kang and Seo [15]. What makes our proof particularly original and elegant is that the rigorous derivation of high-order terms follows almost immediately. The extension of the techniques used in [14], [7], and [26] to construct higher-order terms in the expansion of $u_\epsilon|_{\partial\Omega}$ as $\epsilon \to 0$ seems to be laborious. Furthermore, the general approach developed in this paper could be carried out to obtain more precise asymptotic formulas for the full Maxwell equations and for the equations of linear elasticity than those derived in [3] and [1]. The method of this paper also enables us to extend the asymptotic expansions to the cases of inhomogeneities with Lipschitz boundaries. Previously, the leading order term was derived under the assumption that inhomogeneities are $C^{1,\alpha}$ smooth [14], [7]. We note that our method works as well even when the inhomogeneities have extreme conductivities ($k = 0$ or $k = \infty$).

Let us now explain what makes this asymptotic formula interesting in electrical impedance tomography (EIT). It is well known that the ultimate objective of EIT is to recover, most efficiently and accurately, the conductivity distribution inside a body from measurements of current flows and voltages on the body's surface. The vast and growing literature reflects the many possible applications of EIT, e.g., for medical diagnosis or nondestructive evaluation of materials [6]. In its most general form EIT is severely ill-posed and nonlinear. Taking advantage of the smallness of the inhomogeneities, Cedio-Fengya, Moskow, and Vogelius [7] used the leading order term in the asymptotic expansion of $u_\epsilon|_{\partial\Omega}$ to find the locations z_l, $l = 1, \dots, m$, of the inhomogeneities and certain properties of the domains B_l, $l = 1, \dots, m$ (relative size, orientation). The algorithm proposed in [7] is based on a least-squares algorithm. Ammari, Moskow, and Vogelius [2] also utilized this leading order term to design a variationally based direct reconstruction method. The new idea in [2] is to form the integral of the "measured boundary data" against harmonic test functions and choose the input current g so as to obtain an expression involving the inverse Fourier transform of distributions supported at the locations z_l, $l = 1, \dots, m$. Applying a direct Fourier transform to this data then pins down the locations. This approach is similar to the ideas used by Calderón [5] in his proof of uniqueness of the linearized conductivity problem and later by Sylvester and Uhlmann in their important work [24] on uniqueness of the three-dimensional inverse conductivity problem. Another algorithm that makes use of an asymptotic expansion of the voltage potentials was derived by Brühl, Hanke, and Vogelius [4]. This algorithm is in the spirit of the linear sampling method of Colton and Kirsch [9].

In all of these algorithms, the locations z_l, $l = 1, \dots, m$, of the inhomogeneities are found with an error $O(\epsilon)$, and little about the domains B_l can be reconstructed. Making use of higher-order terms in the asymptotic expansion of $u_\epsilon|_{\partial\Omega}$, we certainly would be able to reconstruct the small inhomogeneities with higher resolution from boundary information about specific solutions to (1.2). Perhaps, more importantly, this would allow us to identify quite general conductivity inhomogeneities without restrictions on their sizes.

The use of higher-order terms in the asymptotic expansion of $u_\epsilon|_{\partial\Omega}$ may also be decisive in dramatically improving the algorithm of Kwon, Seo, and Yoon [19], which is based on the observation of the pattern of a simple weighted combination of an input current g of the form $g = a \cdot \nu$ for some constant vector a and the corresponding output voltage. We also believe that the use of such higher-order terms would improve

the algorithm of Mast, Nachman, and Waag [20], which uses eigenfunctions of the scattering operator.

This paper is organized as follows. In section 2, we collect some notation and preliminary results regarding layer potentials. In section 3, we introduce the GPTs associated with the domains D_l and the conductivities k_l. In section 4, we provide a rigorous derivation of high-order terms in the asymptotic expansion of the output voltage potentials. For reasons of brevity we restrict a significant part of this derivation to the case of a single inhomogeneity ($m = 1$). The proof in the case of multiple well-separated inhomogeneities may be derived by a fairly straightforward iteration of the arguments we present; however, we leave the details to the reader.

2. Layer potentials for the Laplacian. Let us first review some well-known properties of the layer potentials for the Laplacian and prove some useful identities.

The theory of layer potentials has been developed in relation to the boundary value problems. Let D be a bounded domain in $\mathbb{R}^d, d \geq 2$. We assume that ∂D is Lipschitz. Let $\Gamma(x)$ be the fundamental solution of the Laplacian Δ,

$$\Gamma(x) = \begin{cases} \dfrac{1}{2\pi} \ln |x|, & d = 2, \\ \dfrac{1}{(2-d)\omega_d} |x|^{2-d}, & d \geq 3, \end{cases} \tag{2.1}$$

where ω_d is the area of $(d-1)$-dimensional unit sphere. The single and double layer potentials of the density function ϕ on D are defined by

$$\mathcal{S}_D\phi(x) := \int_{\partial D} \Gamma(x-y)\phi(y)d\sigma(y), \quad x \in \mathbb{R}^d, \tag{2.2}$$

$$\mathcal{D}_D\phi(x) := \int_{\partial D} \frac{\partial}{\partial \nu_y}\Gamma(x-y)\phi(y)d\sigma(y), \quad x \in \mathbb{R}^d \setminus \partial D. \tag{2.3}$$

For a function u defined on $\mathbb{R}^d \setminus \partial D$, we denote

$$\frac{\partial}{\partial \nu^{\pm}} u(x) := \lim_{t \to 0^+} \langle \nabla u(x \pm t\nu_x), \nu_x \rangle, \quad x \in \partial D,$$

if the limit exists. Here ν_x is the outward unit normal to ∂D at x.

The proof of the following trace formula can be found in [11], [13], [21] (for Lipschitz domains, see [25]):

$$\frac{\partial}{\partial \nu^{\pm}} \mathcal{S}_D\phi(x) = \left(\pm\frac{1}{2}I + \mathcal{K}_D^*\right)\phi(x), \tag{2.4}$$

$$(\mathcal{D}_D\phi)|_{\pm} = \left(\mp\frac{1}{2}I + \mathcal{K}_D\right)\phi(x), \quad x \in \partial D, \tag{2.5}$$

where

$$\mathcal{K}_D\phi(x) = \frac{1}{\omega_d} \text{p.v.} \int_{\partial D} \frac{\langle x-y, \nu_y \rangle}{|x-y|^d}\phi(y)d\sigma(y)$$

and $\mathcal{K}_D^*$ is the L^2-adjoint of $\mathcal{K}_D$. When ∂D is Lipschitz, $\mathcal{K}_D$ is a singular integral operator and known to be bounded on $L^2(\partial\Omega)$ [8]. Let $L_0^2(\partial D) := \{f \in L^2(\partial D) :$

$\int_{\partial D} f d\sigma = 0\}$. The following results are due to Verchota [25] and Escauriaza, Fabes, and Verchota [10].

THEOREM 2.1 (see [10], [25]). *$\lambda I - \mathcal{K}_D^*$ is invertible on $L_0^2(\partial D)$ if $|\lambda| \geq \frac{1}{2}$, and for $\lambda \in (-\infty, -\frac{1}{2}] \cup (\frac{1}{2}, \infty)$, $\lambda I - \mathcal{K}_D^*$ is invertible on $L^2(\partial D)$.*

For proofs when ∂D is smooth, see [11], [13].

The following theorem was proved in [15], [16], [17].

THEOREM 2.2. *Suppose that D is a domain compactly contained in Ω with a connected Lipschitz boundary and conductivity k. Then the solution u of the problem*

$$\begin{cases} \nabla \cdot ((1 + (k-1)\chi(D))\nabla u) = 0 \quad \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu}\Big|_{\partial\Omega} = g \end{cases} \tag{2.6}$$

is represented as

$$u(x) = H(x) + \mathcal{S}_D\phi(x), \quad x \in \Omega, \tag{2.7}$$

where the harmonic function H is given by

$$H(x) = -\mathcal{S}_\Omega(g)(x) + \mathcal{D}_\Omega(f)(x), \quad x \in \Omega, \quad f := u|_{\partial\Omega}, \tag{2.8}$$

and $\phi \in L_0^2(\partial D)$ satisfies the integral equation

$$\left(\frac{k+1}{2(k-1)} I - \mathcal{K}_D^*\right)\phi = \frac{\partial H}{\partial \nu}\Big|_{\partial D} \quad \text{on } \partial D. \tag{2.9}$$

Moreover, $\forall\, n \in \mathbb{N}$, there exists a constant $C_n = C(n, \Omega, dist(D, \partial\Omega))$ independent of $|D|$ and k such that

$$\|H\|_{\mathcal{C}^n(\overline{D})} \leq C_n \|g\|_{L^2(\partial\Omega)}. \tag{2.10}$$

Proof. The representation formula (2.7) was proved in [15], [17]. Equation (2.10) was proved in [16] for $d = 2$, and it is easily seen that the same proof works for $d = 3$. We only need to check carefully whether the constant C_n in the estimate (2.10) is independent of $|D|$. Before doing this, let us point out that the harmonic function H can be computed explicitly from the boundary measurements $(\frac{\partial u}{\partial \nu}\big|_{\partial\Omega}, u\big|_{\partial\Omega})$, and the density ϕ is uniquely and explicitly determined by the domain D and the harmonic function H. The decomposition of the function u into a harmonic part H and a refraction part $\mathcal{S}_D\phi$ is unique [15], [17]. The representation formula (2.7) seems to inherit geometric properties of D.

Suppose that $\text{dist}(D, \partial\Omega) > 2c_0$ for some constant $c_0 > 0$. From the definition of H in (2.8) it is easy to see that

$$\|H\|_{\mathcal{C}^n(\overline{D})} \leq C_n\Big(\|g\|_{L^2(\partial\Omega)} + \|u|_{\partial\Omega}\|_{L^2(\partial\Omega)}\Big), \tag{2.11}$$

where C_n depends only on n, $\partial\Omega$, and c_0. Let $\vec{\alpha}$ be a vector field supported in the set $\text{dist}(x, \partial\Omega) < 2c_0$ such that $\vec{\alpha} \cdot \nu(x) \geq \delta$ for some $\delta > 0\ \forall\ x \in \partial\Omega$. Using the Rellich identity with this $\vec{\alpha}$, we can show that

$$\left\|\frac{\partial u}{\partial T}\right\|_{L^2(\partial\Omega)} \leq C\Big(\|g\|_{L^2(\partial\Omega)} + \|\nabla u\|_{L^2(\Omega\setminus\overline{D})}\Big),$$

where C depends only on $\partial\Omega$ and c_0 and $T(x)$ is the tangent vector to $\partial\Omega$ at x. (See the proof of Lemma 2.1 of [12] for details of the proof.) Observe that

$$\begin{aligned}\|\nabla u\|^2_{L^2(\Omega\setminus\overline{D})} &\le C\int_\Omega (1+(k-1)\chi(D))\nabla u\cdot\nabla u\,dx\\ &= C\int_{\partial\Omega} gu\,d\sigma\\ &\le C\|g\|_{L^2(\partial\Omega)}\|u|_{\partial\Omega}\|_{L^2(\partial\Omega)}.\end{aligned}$$

Since $\int_{\partial\Omega} u\,d\sigma = 0$, it follows from the Poincaré inequality that

$$\|u|_{\partial\Omega}\|_{L^2(\partial\Omega)} \le C\left\|\frac{\partial u}{\partial T}\right\|_{L^2(\partial\Omega)}.$$

Thus we obtain

$$\|u|_{\partial\Omega}\|^2_{L^2(\partial\Omega)} \le C\Big(\|g\|^2_{L^2(\partial\Omega)} + \|g\|_{L^2(\partial\Omega)}\|u|_{\partial\Omega}\|_{L^2(\partial\Omega)}\Big),$$

and hence

$$\|u|_{\partial\Omega}\|_{L^2(\partial\Omega)} \le C\|g\|_{L^2(\partial\Omega)}.$$

From (2.11) we finally obtain (2.10). □

Using the above representation we can derive a formula similar to (1.4) which is potentially useful in detecting the inhomogeneities (see the remark at the end of this paper). However, it uses the function H, which depends on D and hence on ϵ. Thus in order to derive (1.4) we will transform it to representations using only background potentials.

Let $N(x,z)$ be the Neumann function for Δ in Ω corresponding to a Dirac mass at z. That is, N is the solution to

$$\begin{cases}\Delta_x N(x,z) = -\delta_z & \text{in } \Omega,\\ \dfrac{\partial N}{\partial\nu}\Big|_{\partial\Omega} = -\dfrac{1}{|\partial\Omega|}.\end{cases} \tag{2.12}$$

In addition, we assume that

$$\int_{\partial\Omega} N(x,y)d\sigma(x) = 0 \quad \text{for } y\in\Omega. \tag{2.13}$$

Let us fix one more notation: For D, a subset of Ω, let

$$N_D f(x) := \int_{\partial D} N(x,y)f(y)d\sigma(y).$$

The following lemma relates the fundamental solution with the Neumann function.

LEMMA 2.3. *For $z\in\Omega$ and $x\in\partial\Omega$, let $\Gamma_z(x) := \Gamma(x-z)$ and $N_z(x) := N(x,z)$. Then*

$$\left(-\frac{1}{2}I + \mathcal{K}_\Omega\right)(N_z)(x) = \Gamma_z(x) \quad \textit{modulo constants}, \quad x\in\partial\Omega, \tag{2.14}$$

or, to be more precise, for any simply connected Lipschitz domain D compactly contained in Ω and for any $g \in L^2_0(\partial D)$, we have

$$\int_{\partial D}\left(-\frac{1}{2}I+\mathcal{K}_\Omega\right)(N_z)(x)g(z)d\sigma(z)=\int_{\partial D}\Gamma_z(x)g(z)d\sigma(z)\quad\forall x\in\partial\Omega. \tag{2.15}$$

Proof. Let $f \in L^2_0(\partial\Omega)$ and define

$$u(z):=\left\langle\left(-\frac{1}{2}I+\mathcal{K}_\Omega\right)(N_z),f\right\rangle_{\partial\Omega},\quad z\in\Omega.$$

Then

$$u(z)=\int_{\partial\Omega}N(x,z)\left(-\frac{1}{2}I+\mathcal{K}^*_\Omega\right)f(x)d\sigma(x).$$

Therefore, $\Delta u = 0$ in Ω and $\frac{\partial u}{\partial\nu}|_{\partial\Omega} = (-\frac{1}{2}I+\mathcal{K}^*_\Omega)f$. Thus by (2.4) we have

$$u(z)-\mathcal{S}_\Omega f(z)=\text{constant},\quad z\in\Omega.$$

Thus if $g \in L^2_0(\partial\Omega)$, then we obtain

$$\begin{aligned}&\int_{\partial\Omega}\int_{\partial D}\left(-\frac{1}{2}I+\mathcal{K}_\Omega\right)(N_z)(x)g(z)d\sigma(z)f(x)d\sigma(x)\\&=\int_{\partial\Omega}\int_{\partial D}\Gamma_z(x)g(z)d\sigma(z)f(x)d\sigma(x).\end{aligned}$$

Since f is arbitrary, we have equation (2.14) or, equivalently, (2.15). This completes the proof. ∎

Let $g \in L^2_0(\partial\Omega)$. Let $U(y) := \int_{\partial\Omega} N(x,y)g(x)d\sigma(x)$. Then U satisfies

$$\begin{cases}\Delta U=0\quad\text{in }\Omega,\\ \dfrac{\partial U}{\partial\nu}|_{\partial\Omega}=g\in L^2_0(\partial\Omega),\\ \displaystyle\int_{\partial\Omega}U(x)d\sigma(x)=0.\end{cases} \tag{2.16}$$

THEOREM 2.4. *The solution u of* (2.6) *can be represented as*

$$u(x)=U(x)-N_D\phi(x),\quad x\in\partial\Omega, \tag{2.17}$$

where ϕ is defined in (2.9).

Proof. By substituting (2.7) into (2.8), we obtain

$$H(x)=-\mathcal{S}_\Omega(g)(x)+\mathcal{D}_\Omega(H|_{\partial\Omega}+(\mathcal{S}_D\phi)|_{\partial\Omega})(x),\quad x\in\Omega.$$

It then follows from (2.5) that

$$\left(\frac{1}{2}I-\mathcal{K}_\Omega\right)(H|_{\partial\Omega})=-(\mathcal{S}_\Omega g)|_{\partial\Omega}+\left(\frac{1}{2}I+\mathcal{K}_\Omega\right)((\mathcal{S}_D\phi)|_{\partial\Omega})\quad\text{on }\partial\Omega. \tag{2.18}$$

Since by Green's theorem $U = -\mathcal{S}_\Omega(g) + \mathcal{D}_\Omega(U|_{\partial\Omega})$ in Ω, we have

$$\left(\frac{1}{2}I - \mathcal{K}_\Omega\right)(U|_{\partial\Omega}) = -(\mathcal{S}_\Omega g)|_{\partial\Omega}. \tag{2.19}$$

Since $\phi \in L_0^2(\partial D)$, it follows from (2.14) that

$$-\left(\frac{1}{2}I - \mathcal{K}_\Omega\right)((N_D\phi)|_{\partial\Omega}) = (\mathcal{S}_D\phi)|_{\partial\Omega}. \tag{2.20}$$

From (2.18), (2.19), and (2.20), we conclude that

$$\left(\frac{1}{2}I - \mathcal{K}_\Omega\right)\left(H|_{\partial\Omega} - U|_{\partial\Omega} + \left(\frac{1}{2}I + \mathcal{K}_\Omega\right)((N_D\phi)|_{\partial\Omega})\right) = 0.$$

Therefore, we have

$$H|_{\partial\Omega} - U|_{\partial\Omega} + \left(\frac{1}{2}I + \mathcal{K}_\Omega\right)((N_D\phi)|_{\partial\Omega}) = C \text{ (constant)}. \tag{2.21}$$

Note that $(\frac{1}{2}I+\mathcal{K}_\Omega)((N_D\phi)|_{\partial\Omega}) = (N_D\phi)|_{\partial\Omega}+(\mathcal{S}_D\phi)|_{\partial\Omega}$. Thus we get from (2.7) and (2.21)

$$u|_{\partial\Omega} = U|_{\partial\Omega} - (N_D\phi)|_{\partial\Omega} + C. \tag{2.22}$$

Since all the functions entering (2.22) belong to $L_0^2(\partial\Omega)$, we conclude that $C = 0$, and the theorem is proved. □

We have a similar representation for solutions of the Dirichlet problem. Let $G(x, y)$ be the Green function for the Dirichlet problem; i.e., the function V defined by $V(x) := \int_{\partial\Omega} \frac{\partial G}{\partial\nu(y)}(x, y)f(y)d\sigma(y)$ is the solution of the problem $\Delta V = 0$ in Ω and $V|_{\partial\Omega} = f$ for any $f \in L^2(\partial\Omega)$. Then we have the following representation theorem.

THEOREM 2.5.

$$\left(\frac{1}{2}I + \mathcal{K}_\Omega^*\right)^{-1}\left(\frac{\partial\Gamma_z(y)}{\partial\nu(y)}\right)(x) = \frac{\partial G_z}{\partial\nu(x)}(x), \quad x \in \partial\Omega,\ z \in \Omega. \tag{2.23}$$

Let u be the solution of (2.6) *with the Neumann condition replaced by the Dirichlet condition $u|_{\partial\Omega} = f$. Then u can be represented as*

$$\frac{\partial u}{\partial\nu}(x) = \frac{\partial V}{\partial\nu}(x) - G_D\phi(x), \quad x \in \partial\Omega, \tag{2.24}$$

where ϕ is defined in (2.9) *and $G_D\phi(x) := \int_{\partial D} \frac{\partial G}{\partial\nu(y)}(x, y)\phi(y)d\sigma(y)$.*

Theorem 2.5 can be proved in the same way as Theorem 2.4. In fact, it is simpler because of the solvability of the Dirichlet problem or, equivalently, the invertibility of $(\frac{1}{2}I + \mathcal{K}_\Omega^*)$. So we omit the proof.

3. Generalized polarization tensors. In this section we introduce the generalized polarization tensors (GPTs) associated with a domain B and a conductivity k. These GPTs are the basic building block for the asymptotic expansions in this paper.

Let B be a Lipschitz bounded domain in $\mathbb{R}^d$ and let the conductivity of B be k $(k \neq 1)$. The polarization tensor $M = (m_{ij})$, $1 \leq i, j \leq d$, is defined by

$$m_{ij} := \left(1 - \frac{1}{k}\right)\left[\delta_{ij}|B| + (k-1)\int_{\partial B} y_i \frac{\partial}{\partial\nu^+}\psi_j(y)d\sigma(y)\right],$$

where ψ_j is the unique solution of the following transmission problem:

$$\begin{cases} \Delta\psi_j(x) = 0, \quad x \in B \cup \mathbb{R}^d \setminus \overline{B}, \\ \psi_j|_+ - \psi_j|_- = 0 \quad \text{on } \partial B, \\ \dfrac{\partial}{\partial \nu^+}\psi_j - k\dfrac{\partial}{\partial \nu^-}\psi_j = \nu_j \quad \text{on } \partial B, \\ \psi_j(x) \to 0 \text{ as } |x| \to \infty. \end{cases}$$

See [23], [7], and [14]. One can easily check, using (2.4), that

$$(k-1)\psi_j = \mathcal{S}_B(\lambda I - \mathcal{K}_B^*)^{-1}(\nu_j).$$

Using (2.4) again, we have

$$\begin{aligned} (k-1)\int_{\partial B} y_i \frac{\partial}{\partial \nu^+}\psi_j(y)d\sigma(y) &= \int_{\partial B} y_i\left(\frac{1}{2}I + \mathcal{K}_B^*\right)(\lambda I - \mathcal{K}_B^*)^{-1}(\nu_j)(y)d\sigma(y) \\ &= -\int_{\partial B} y_i\nu_j d\sigma(y) + \left(\lambda + \frac{1}{2}\right)\int_{\partial B} y_i(\lambda I - \mathcal{K}_B^*)^{-1}(\nu_j)(y)d\sigma(y) \\ &= -\delta_{ij}|B| + \frac{k}{k-1}\int_{\partial B} y_i(\lambda I - \mathcal{K}_B^*)^{-1}(\nu_j)(y)d\sigma(y). \end{aligned}$$

Therefore we prove that the polarization tensor M associated with B and k is given by

$$m_{ij} = \int_{\partial B} y_i(\lambda I - \mathcal{K}_B^*)^{-1}(\nu_j)(y)d\sigma(y). \tag{3.1}$$

Recall $\lambda := \frac{k+1}{2(k-1)}$.

For a multi-index $i = (i_1, \dots, i_d) \in \mathbb{N}^d$, let $\partial^i f = \partial_1^{i_1} \cdots \partial_d^{i_d} f$ and $x^i := x_1^{i_1} \cdots x_d^{i_d}$. For $i, j \in \mathbb{N}^d$, we define the *GPT* M_{ij} by

$$M_{ij} := \int_{\partial B} y^j \phi_i(y)d\sigma(y), \tag{3.2}$$

where ϕ_i is defined by

$$\phi_i(x) := (\lambda I - \mathcal{K}_B^*)^{-1}\left(\frac{1}{i!}\nu_y \cdot \nabla y^i\right)(x), \quad x \in \partial B.$$

4. Derivation of the full asymptotic formula. In this section we derive our asymptotic formula (1.4). As stated in the introduction, we restrict our derivation to the case of a single inhomogeneity ($m = 1$). We only give the details when considering the difference between the fields corresponding to one and zero inhomogeneities. In order to further simplify notation we assume that the single inhomogeneity D has the form $D = \epsilon B + z$, where $z \in \Omega$ and B is a bounded Lipschitz domain in $\mathbb{R}^d$ containing the origin. Suppose that the conductivity of D is k. Let $\lambda := \frac{k+1}{2(k-1)}$. Then by (2.7) and (2.9), the solution u of (2.6) takes the form

$$u(x) = U(x) - N_D(\lambda I - \mathcal{K}_D^*)^{-1}\left(\frac{\partial H}{\partial \nu}|_{\partial D}\right)(x), \quad x \in \partial\Omega,$$

where U is the background potential given in (2.16).

Define

$$H_n(x) := \sum_{|i|=0}^{n} \frac{1}{i!}(\partial^i H)(z)(x-z)^i.$$

Here we use the multi-index notation $i = (i_1, \dots, i_d) \in \mathbb{N}^d$. Then we have from (2.10) that

$$\begin{aligned}\left\| \frac{\partial H}{\partial \nu} - \frac{\partial H_n}{\partial \nu} \right\|_{L^2(\partial D)} &\le \sup_{x \in \partial D} |\nabla H(x) - \nabla H_n(x)| |\partial D|^{1/2} \\ &\le \|H\|_{\mathcal{C}^{n+1}(\overline{D})} |x-z|^n |\partial D|^{1/2} \\ &\le C \|g\|_{L^2(\partial\Omega)} \epsilon^n |\partial D|^{1/2}.\end{aligned}$$

Note that

$$\text{if } \int_{\partial D} h d\sigma = 0, \text{ then } \int_{\partial D} (\lambda I - \mathcal{K}_D^*)^{-1} h d\sigma = 0. \tag{4.1}$$

If $\int_{\partial D} h d\sigma = 0$, then we have for $x \in \partial\Omega$ that

$$\begin{aligned}\left| N_D(\lambda I - \mathcal{K}_D^*)^{-1} h(x) \right| &= \left| \int_{\partial D} [N(x-y) - N(x-z)](\lambda I - \mathcal{K}_D^*)^{-1} h(y) d\sigma(y) \right| \\ &\le C\epsilon |\partial D|^{1/2} \|h\|_{L^2(\partial D)}.\end{aligned}$$

It then follows that

$$\begin{aligned}\sup_{x \in \partial D} \left| N_D(\lambda I - \mathcal{K}_D^*)^{-1} \left(\frac{\partial H}{\partial \nu}|_{\partial D} - \frac{\partial H_n}{\partial \nu}|_{\partial D} \right)(x) \right| &\le C\epsilon |\partial D|^{1/2} \left\| \frac{\partial H}{\partial \nu} - \frac{\partial H_n}{\partial \nu} \right\|_{L^2(\partial D)} \\ &\le C \|g\|_{L^2(\partial\Omega)} \epsilon^{d+n}.\end{aligned}$$

Therefore, we have

$$u(x) = U(x) - N_D(\lambda I - \mathcal{K}_D^*)^{-1} \left(\frac{\partial H_n}{\partial \nu}|_{\partial D} \right)(x) + O(\epsilon^{d+n}), \quad x \in \partial\Omega, \tag{4.2}$$

where the $O(\epsilon^{d+n})$ term is dominated by $C\|g\|_{L^2(\partial\Omega)}\epsilon^{d+n}$ for some C depending only on c_0. Note that

$$(\lambda I - \mathcal{K}_D^*)^{-1} \left(\frac{\partial H_n}{\partial \nu}|_{\partial D} \right)(x) = \sum_{|i|=1}^{n} (\partial^i H)(z) (\lambda I - \mathcal{K}_D^*)^{-1} \left(\frac{1}{i!} \nu_x \cdot \nabla (x-z)^i \right)(x).$$

Since $D = \epsilon B + z$, one can prove by using the change of variables $y = \frac{x-z}{\epsilon}$ and the expression of $\mathcal{K}_D^*$ defined as the L^2-adjoint of $\mathcal{K}_D$ that

$$(\lambda I - \mathcal{K}_D^*)^{-1} \left(\frac{1}{i!} \nu_x \cdot \nabla (x-z)^i \right)(x) = \epsilon^{|i|-1} (\lambda I - \mathcal{K}_B^*)^{-1} \left(\frac{1}{i!} \nu_y \cdot \nabla y^i \right) \left(\frac{1}{\epsilon}(x-z) \right).$$

Put

$$\phi_i(x) := (\lambda I - \mathcal{K}_B^*)^{-1} \left(\frac{1}{i!} \nu_y \cdot \nabla y^i \right)(x), \quad x \in \partial B. \tag{4.3}$$

Then we get

$$
\begin{aligned}
(4.4)\qquad & N_D(\lambda I - \mathcal{K}_D^*)^{-1}\left(\frac{\partial H_n}{\partial \nu}|_{\partial D}\right)(x) \\
&= \sum_{|i|=1}^{n} (\partial^i H)(z)\epsilon^{|i|-1} \int_{\partial D} N(x,y)\phi_i(\epsilon^{-1}(y-z))d\sigma(y) \\
&= \sum_{|i|=1}^{n} (\partial^i H)(z)\epsilon^{|i|+d-2} \int_{\partial B} N(x,\epsilon y+z)\phi_i(y)d\sigma(y).
\end{aligned}
$$

We now expand $N(x,\epsilon y+z)$ asymptotically as $\epsilon \to 0$. By (2.14) we have the following relation:

$$
\left(-\frac{1}{2}I + \mathcal{K}_\Omega\right)[N(\cdot,\epsilon y+z)](x) = \Gamma(x-z-\epsilon y) \quad \text{modulo constants}, \quad x \in \partial\Omega.
$$

Using the Taylor expansion

$$
\Gamma(x-\epsilon y) = \sum_{|j|=0}^{+\infty} \frac{(-1)^j}{j!}\epsilon^{|j|}\partial^j(\Gamma(x))y^j,
$$

we obtain

$$
\begin{aligned}
\left(-\frac{1}{2}I + \mathcal{K}_\Omega\right)[N(\cdot,\epsilon y+z)](x) &= \sum_{|j|=0}^{+\infty} \frac{(-1)^j}{j!}\epsilon^{|j|}\partial^j(\Gamma(x-z))y^j \\
&= \sum_{|j|=0}^{+\infty} \frac{(-1)^j}{j!}\epsilon^{|j|}\partial_x^j\left(\left(-\frac{1}{2}I + \mathcal{K}_\Omega\right)N(\cdot,z)(x)\right)y^j \\
&= \sum_{|j|=0}^{+\infty} \frac{1}{j!}\epsilon^{|j|}\left(\left(-\frac{1}{2}I + \mathcal{K}_\Omega\right)\partial_z^j N(\cdot,z)(x)\right)y^j \\
&= \left(-\frac{1}{2}I + \mathcal{K}_\Omega\right)\left[\sum_{|j|=0}^{+\infty} \frac{1}{j!}\epsilon^{|j|}\partial_z^j N(\cdot,z)y^j\right](x).
\end{aligned}
$$

Since $\int_{\partial\Omega} N(x,w)d\sigma(x) = 0 \ \forall\ w \in \Omega$, we have the following asymptotic expansion of the Neumann function, which is of independent interest.

LEMMA 4.1. *For $x \in \partial\Omega$, $z \in \Omega$, and $y \in \partial B$, and as $\epsilon \to 0$,*

$$
(4.5)\qquad N(x,\epsilon y+z) = \sum_{|j|=0}^{+\infty} \frac{1}{j!}\epsilon^{|j|}\partial_z^j N(x,z)y^j.
$$

We now have from (4.4)

$$
\begin{aligned}
& N_D(\lambda I - \mathcal{K}_D^*)^{-1}\left(\frac{\partial H_n}{\partial \nu}|_{\partial D}\right)(x) \\
&= \sum_{|i|=1}^{n} (\partial^i H)(z)\epsilon^{|i|+d-2} \sum_{|j|=0}^{+\infty} \frac{1}{j!}\epsilon^{|j|}\partial_z^j N(x,z) \int_{\partial B} y^j \phi_i(y)d\sigma(y).
\end{aligned}
$$

Observe that since H is a harmonic function in Ω we may compute

$$\sum_{|i|=l} \frac{1}{i!}(\partial^i H)(z)\Delta(y^i) = \Delta_y \left(\sum_{|i|=l} \frac{1}{i!}(\partial^i H)(z) y^i \right) = 0,$$

and therefore, by Green's theorem, it follows that

$$\int_{\partial B} \sum_{|i|=l} \frac{1}{i!}(\partial^i H)(z)\nabla(y^i)\cdot \nu(y)\, d\sigma(y) = 0.$$

Thus, in view of (4.3), the following identity holds by using observation (4.1):

$$\sum_{|i|=l} (\partial^i H)(z) \int_{\partial B} \phi_i(y) d\sigma(y) = 0 \quad \forall\, l \geq 1. \tag{4.6}$$

In fact, this follows immediately from (4.1). Recall now that

$$\epsilon^{d-2} N(x, \epsilon y + z) = \epsilon^{d-2} \sum_{|j|=0}^{n-|i|+1} \frac{1}{j!} \epsilon^{|j|} \partial_z^j N(x,z) y^j + O(\epsilon^{d+n-|i|}) \quad \forall\, i, 1 \leq |i| \leq n,$$

and on the other hand $M_{ij} = \int_{\partial B} y^j \phi_i(y) d\sigma(y)$ is the GPT associated with the domain B and the conductivity k to obtain the following pointwise asymptotic formula: For $x \in \partial\Omega$,

$$u(x) = U(x) - \epsilon^{d-2} \sum_{|i|=1}^{n} \sum_{|j|=1}^{n-|i|+1} \frac{1}{j!} \epsilon^{|i|+|j|} (\partial^i H)(z) M_{ij} \partial_z^j N(x,z) + O(\epsilon^{d+n}). \tag{4.7}$$

Observing that the formula (4.7) still contains $\partial^i H$ factors, we see that the remaining task is to convert (4.7) to a formula given solely by U and its derivatives.

As a simplest case, let us now take $n = 1$ to find the leading order term in the asymptotic expansion of $u|_{\partial\Omega}$ as $\epsilon \to 0$. From (2.7) and (2.17), we get

$$\|H - U\|_{L^\infty(\partial\Omega)} \leq C\epsilon^{\frac{d}{2}} \|\phi\|_{L^2(\partial D)} \leq C\epsilon^{\frac{d}{2}} \|g\|_{L^2(\partial\Omega)}$$

for some C depending only on Ω and c_0. It then follows from the maximum principle that

$$\|H - U\|_{L^\infty(\Omega)} \leq C\epsilon^{\frac{d}{2}} \|g\|_{L^2(\partial\Omega)}.$$

Then, from the mean value property of harmonic functions, we obtain

$$|\nabla H(z) - \nabla U(z)| \leq C\epsilon^{\frac{d}{2}} \|g\|_{L^2(\partial\Omega)}.$$

It thus follows from (4.7) that

$$u(x) = U(x) - \epsilon^d \sum_{|i|=1,|j|=1} (\partial^i U)(z) M_{ij} \partial^j N(x,z) + O(\epsilon^{d+1}), \quad x \in \partial\Omega, \tag{4.8}$$

which is, in view of (3.1), exactly the formula derived in [14] and [7] when D has $C^{1,\alpha}$ boundary.

We now return to (4.7). Recalling that by Green's theorem $U = -\mathcal{S}_\Omega(g) + \mathcal{D}_\Omega(U|_{\partial\Omega})$ in Ω, substitution of (4.7) into (2.8) immediately yields that, for any $x \in \Omega$,

(4.9)
$$H(x) = U(x) - \epsilon^{d-2} \sum_{|i|=1}^{n} \sum_{|j|=1}^{n-|i|+1} \frac{1}{j!} \epsilon^{|i|+|j|} (\partial^i H)(z) M_{ij} \mathcal{D}_\Omega(\partial_z^j N(\cdot, z))(x) + O(\epsilon^{d+n}).$$

In (4.9) the remainder $O(\epsilon^{d+n})$ is uniform in the $\mathcal{C}^n$ norm on any compact subset of Ω for any n, and therefore

$$(\partial^l H)(z) + \sum_{|i|=1}^{n} \epsilon^{d-2} \sum_{|j|=1}^{n-|i|+1} \epsilon^{|i|+|j|} (\partial^i H)(z) P_{ijl} = (\partial^l U)(z) + O(\epsilon^{d+n}) \tag{4.10}$$

$\forall\, l \in \mathbb{N}^d$ with $|l| \le n$, where

$$P_{ijl} = \frac{1}{j!} M_{ij} \partial_x^l \mathcal{D}_\Omega(\partial_z^j N(\cdot, z))|_{x=z}. \tag{4.11}$$

Define the operator

$$\mathcal{P}_\epsilon : (v_l)_{l \in \mathbb{N}^d, |l| \le n} \mapsto \left(v_l + \epsilon^{d-2} \sum_{|i|=1}^{n} \sum_{|j|=1}^{n-|i|+1} \epsilon^{|i|+|j|} v_i P_{ijl} \right)_{l \in \mathbb{N}^d, |l| \le n}.$$

Observe that

$$\mathcal{P}_\epsilon = I + \epsilon^d \mathcal{R}_1 + \cdots + \epsilon^{n+d-1} \mathcal{R}_{n-1}.$$

Defining the matrices $\mathcal{Q}_p, p = 1, \ldots, n-1$, by

(4.12)
$$(I + \epsilon^d \mathcal{R}_1 + \cdots + \epsilon^{n+d-1} \mathcal{R}_{n-1})^{-1} = I + \epsilon^d \mathcal{Q}_1 + \cdots + \epsilon^{n+d-1} \mathcal{Q}_{n-1} + O(\epsilon^{n+d})$$

for small ϵ, we finally obtain that

$$((\partial^i H)(z))_{i \in \mathbb{N}^d, |i| \le n} = \left(I + \sum_{p=1}^{n} \epsilon^{d+p-1} \mathcal{Q}_p \right) ((\partial^i U)(z))_{i \in \mathbb{N}^d, |i| \le n} + O(\epsilon^{d+n}), \tag{4.13}$$

which yields the main result of this paper stated in Theorem 1.1.

We also have a complete asymptotic expansion of the solutions of the Dirichlet problem.

THEOREM 4.2. *Suppose that the inhomogeneity consists of a single component, and let u be the solution of* (1.2) *with the Neumann condition replaced by the Dirichlet condition $u|_{\partial\Omega} = f$. Let V be the solution of $\Delta V = 0$ in Ω with $V|_{\partial\Omega} = f$. The following pointwise asymptotic expansion on $\partial\Omega$ holds for $d = 2, 3$:*

$$\begin{aligned} \frac{\partial u}{\partial \nu}(x) = {} & \frac{\partial V}{\partial \nu}(x) - \epsilon^{d-2} \sum_{|i|=1}^{n} \sum_{|j|=1}^{n-|i|+1} \frac{1}{j!} \epsilon^{|i|+|j|} \\ & \times \left[\left(\left(I + \sum_{p=1}^{n+2-|i|-|j|-d} \epsilon^{d+p-1} \mathcal{Q}_p \right) (\partial^l V(z)) \right)_i M_{ij} \partial_z^j \frac{\partial}{\partial \nu_x} G(x, z) \right] \\ & + O(\epsilon^{d+n}), \end{aligned} \tag{4.14}$$

where the remainders $O(\epsilon^{d+n})$ *are dominated by* $C\epsilon^{d+n}\|f\|_{H^{1/2}(\partial\Omega)}$ *for some* C *independent of* $x \in \partial\Omega$. *Here* $G(x,z)$ *is the Dirichlet Green function,* M_{ij}, $i,j \in \mathbb{N}^d$, *are the GPTs, and* $\mathcal{Q}_p$ *is the operator defined in* (4.12), *where* $\mathcal{P}_{ijk}$ *is defined, in this case, by*

$$P_{ijl} = \frac{1}{j!} M_{ij} \partial_x^l \mathcal{S}_\Omega \left(\partial_z^j \left(\frac{\partial}{\partial \nu_x} G \right)(\cdot, z) \right)|_{x=z}. \tag{4.15}$$

Theorem 4.2 can be proved in the exactly same manner as Theorem 1.1. We begin with Theorem 2.5. Then the same arguments give us

$$u(x) = V(x) - \epsilon^{d-2} \sum_{|i|=1}^{n} \sum_{|j|=1}^{n-|i|+1} \frac{1}{j!} \epsilon^{|i|+|j|} (\partial^i H)(z) M_{ij} \partial_z^j G(x,z) + O(\epsilon^{d+n}).$$

From this we can get (4.14) as before.

We conclude this paper by making a remark. The following formula is not exactly an asymptotic formula. However, since the formula is simple and has some potential applicability in solving the inverse conductivity problem, we make a record of it as a theorem.

THEOREM 4.3. *We have*

$$u(x) = H(x) + \epsilon^{d-2} \sum_{|i|=1}^{n} \sum_{|j|=1}^{n-|i|+1} \frac{1}{j!} \epsilon^{|i|+|j|} \partial^i H(z) M_{ij} \partial^j \Gamma(x-z) + O(\epsilon^{d+n}), \tag{4.16}$$

where $x \in \Omega_0$ *and the* $O(\epsilon^{d+n})$ *term is dominated by* $C\|g\|_{L^2(\partial\Omega)}\epsilon^{d+n}$ *for some* C *depending only on* c_0, *and* H *is given in* (2.8).

Acknowledgments. The authors would like to thank Gunther Uhlmann for his invitation and the Mathematical Sciences Research Institute for partial support and for providing a stimulating environment.

REFERENCES

[1] C. ALVES AND H. AMMARI, *Boundary integral formulae for the reconstruction of imperfections of small diameter in an elastic medium*, SIAM J. Appl. Math., 62 (2001), pp. 94–106.

[2] H. AMMARI, S. MOSKOW, AND M. VOGELIUS, *Boundary integral formulas for the reconstruction of electromagnetic imperfections of small diameter*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 49–66.

[3] H. AMMARI, M. VOGELIUS, AND D. VOLKOV, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of imperfections of small diameter* II. *The full Maxwell equations*, J. Math. Pures Appl. (9), 80 (2001), pp. 769–814.

[4] M. BRÜHL, M. HANKE, AND M. VOGELIUS, *A direct impedance tomography algorithm for locating small inhomogeneities*, Numer. Math., 93 (2003), pp. 635–654.

[5] A. P. CALDERÓN, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and Its Applications to Continuum Physics, Soc. Brasileira de Matemática, Rio de Janeiro, 1980, pp. 65–73.

[6] M. CHENEY, D. ISAACSON, AND J.C. NEWELL, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.

[7] D.J. CEDIO-FENGYA, S. MOSKOW, AND M. VOGELIUS, *Identification of conductivity imperfections of small diameter by boundary measurements: Continuous dependence and computational reconstruction*, Inverse Problems, 14 (1998), pp. 553–595.

[8] R.R. COIFMAN, A. MCINTOSH, AND Y. MEYER, *L'integrale de Cauchy definit un opérateur bournée sur* L^2 *pour courbes lipschitziennes*, Ann. of Math. (2), 116 (1982), pp. 361–387.

[9] D. COLTON AND A. KIRSCH, *A simple method for solving inverse scattering problems in the resonance region*, Inverse Problems, 12 (1996), pp. 383–393.

[10] L. Escauriaza, E.B. Fabes, and G. Verchota, *On a regularity theorem for weak solutions to transmission problems with internal Lipschitz boundaries*, Proc. Amer. Math. Soc., 115 (1992), pp. 1069–1076.

[11] E.B. Fabes, M. Jodeit, and N.M. Riviére, *Potential techniques for boundary value problems on $\mathcal{C}^1$ domains*, Acta Math., 141 (1978), pp. 165–186.

[12] E. Fabes, H. Kang, and J.K. Seo, *Inverse conductivity problem with one measurement: Error estimates and approximate identification for perturbed disks*, SIAM J. Math. Anal., 30 (1999), pp. 699–720.

[13] G.B. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

[14] A. Friedman and M. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Ration. Mech. Anal., 105 (1989), pp. 299–326.

[15] H. Kang and J.K. Seo, *Layer potential technique for the inverse conductivity problem*, Inverse Problems, 12 (1996), pp. 267–278.

[16] H. Kang and J.K. Seo, *Identification of domains with near-extreme conductivity: Global stability and error estimates*, Inverse Problems, 15 (1999), pp. 851–867.

[17] H. Kang and J.K. Seo, *Recent progress in the inverse conductivity problem with single measurement*, in Inverse Problems and Related Fields, CRC Press, Boca Raton, FL, 2000, pp. 69–80.

[18] R.E. Kleinman and T.B.A. Senior, *Rayleigh scattering*, in Low and High Frequency Asymptotics, V.K. Varadan and V.V. Varadan, eds., North-Holland, Amsterdam, 1986, pp. 1–70.

[19] O. Kwon, J.K. Seo, and J.R. Yoon, *A real-time algorithm for the location search of discontinuous conductivities with one measurement*, Comm. Pure Appl. Math., 55 (2002), pp. 1–29.

[20] T.D. Mast, A. Nachman, and R.C. Waag, *Focusing and imaging using eigenfunctions of the scattering operator*, J. Acoust. Soc. Amer., 102 (1997), pp. 715–725.

[21] A. Nachmann, *Reconstructions from boundary measurements*, Ann. of Math. (2), 128 (1988), pp. 531–587.

[22] G. Pólya and G. Szegö, *Isoperimetric Inequalities in Mathematical Physics*, Ann. of Math. Stud. 27, Princeton University Press, Princeton, NJ, 1951.

[23] M. Schiffer and G. Szegö, *Virtual mass and polarization*, Trans. Amer. Math. Soc., 67 (1949), pp. 130–205.

[24] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.

[25] G.C. Verchota, *Layer potentials and boundary value problems for Laplace's equation in Lipschitz domains*, J. Funct. Anal., 59 (1984), pp. 572–611.

[26] M. Vogelius and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 723–748.

SIAM J. MATH. ANAL.
Vol. 34, No. 5, pp. 1167–1182

BROWNIAN TRAJECTORY IS A REGULAR LATERAL BOUNDARY FOR THE HEAT EQUATION*

N. V. KRYLOV†

Abstract. The one-dimensional heat equation in the domain $x > w_t$, $t > 0$, is considered. Here w_t is a trajectory of Brownian motion. For almost any trajectory, it is proved that if the boundary data are continuous, then the solution is continuous in the closure of the domain. The proof is based on Davis's law of square root for Brownian motion or on its weaker version, which is obtained by using the theory of stochastic partial differential equations.

Key words. fine properties of Brownian motion, stochastic partial differential equations, heat equation in irregular cylinders

AMS subject classifications. 60G17, 35K05, 60H15

PII. S0036141002402980

1. Introduction and main results.

Let $w_t, t \geq 0$, be a one-dimensional Wiener process on a complete probability space $(\Omega, \mathcal{F}, P)$ and constants $T, \nu \in (0, \infty)$. Define

$$Q = Q(w.) = \{(t, x) : t \in (0, T), x > w_t\}.$$

This is a region depending on ω. For each fixed ω we consider the following boundary value problem:

$$u_t(t,x) + \frac{1}{2}\nu^2 u_{xx}(t,x) = 0, \quad (t,x) \in Q(w.), \tag{1.1}$$

$$u(t, w_t) = g(t, w_t) \quad \text{for} \quad t \in [0, T), \qquad u(T, x) = g(T, x) \quad \text{for} \quad x \geq w_T. \tag{1.2}$$

We assume that g is a bounded continuous function, and by solution u of (1.1)–(1.2) we mean Perron's or probabilistic solution. One of the main goals of this article is to prove the following result, saying that, with probability one, any point of the parabolic boundary of $Q(w.)$ is regular.

THEOREM 1.1. *There is a measurable set $\Omega' \subset \Omega$ such that $P(\Omega') = 1$ and for each $\omega \in \Omega'$, any continuous bounded g, and any $t_0 \in [0, T]$, we have*

$$\lim_{\substack{(t,x) \to (t_0, w_{t_0}) \\ (t,x) \in Q(w.)}} u(t,x) = g(t_0, w_{t_0}). \tag{1.3}$$

Parabolic equations in noncylindrical domains have been considered for quite some time, and many important results are known for them. We refer the interested reader to [1] and the very extensive bibliography in this book. Most of the literature treats the case of boundary which is Hölder (1/2+), and in this case it is possible to get Hölder estimates up to the boundary if g is Hölder continuous. However, a typical Wiener trajectory is only (1/2−) Hölder continuous. Therefore, in this article we only deal with regularity and not with Hölder estimates up to the boundary.

*Received by the editors February 22, 2002; accepted for publication (in revised form) September 20, 2002; published electronically April 15, 2003. The work was partially supported by NSF grant DMS-0140405.

http://www.siam.org/journals/sima/34-5/40298.html

†University of Minnesota, 127 Vincent Hall, Minneapolis, MN 55455 (krylov@math.umn.edu).

The best tractable conditions for the caloric regularity of a point on the lateral boundary are expressed in terms of Khinchin's law of the iterated logarithm or, more generally, Kolmogorov–Petrovskii criteria. However, Lévy's theorem says that (a.s.)

$$\varliminf_{\substack{h=t-s\downarrow 0\\ t,s\leq T}} \frac{w_t-w_s}{\sqrt{2h|\ln h|}}=-1,$$

so that, in particular, $w_{t+h}-w_t\leq -\sqrt{h|\ln h|}$ for appropriate t's in $[0,T]$ and as small $h>0$ as we like. Such functions are way off the range of applicability of Kolmogorov–Petrovskii criteria, and actually only chaotic behavior of Brownian trajectories saves the regularity. In section 2 we give the proof of Theorem 1.1 based on Theorem 1.5, which is a weak version of Davis's law of square root (see [2]).

THEOREM 1.2. *With probability one*

$$\inf_{t\in[0,T]} \varlimsup_{h\downarrow 0} \frac{w_{t+h}-w_t}{\sqrt{h}}=-1. \tag{1.4}$$

The author's interest in Theorem 1.1 arose from the theory of stochastic partial differential equations (SPDEs). To understand how it happened, consider the following situation. Take a number $\sigma\geq 0$ satisfying $\sigma^2<2$ and a nonnegative function $\zeta\in C_0^\infty(0,\infty)$ such that $\zeta(x)=1$ for $x\in[1,2]$. Let $f(t,x)$ be a solution of

$$df(t,x)=f_{xx}(t,x)\,dt+\sigma f_x(t,x)\,dw_t \tag{1.5}$$

for $t\in(0,T)$ and $x>0$ with zero lateral condition and with $f(0,x)=\zeta(x)$ for $x>0$. Here the lateral condition is understood in a certain generalized sense (we say more about it in section 3). Interestingly enough, the general theory of SPDEs developed in [11] implies that the solution f exists for any $\sigma^2<2$, but (a.s.) it is continuous up to the boundary, thus assuming the boundary data, only if $\sigma^2<1$. The author's numerous attempts failed to use the L_p-theory of SPDEs and get *any* information about the usual continuity of f up to the boundary if $1\leq\sigma^2<2$. One may attribute this failure to the fact that in the theory of SPDEs the moments of Hölder constants are estimated and, probably, for $1\leq\sigma^2<2$ the moments are infinite. Anyhow, Theorem 1.1 allows us to prove in section 3 (see Theorem 3.2) the continuity of f in $[0,T]\times[0,\infty)$ assuming that $\sigma^2<2$ and observing that the function $u(t,x):=f(t,\sigma(x-w_t))$ satisfies the equation

$$u_t=\frac{1}{2}\nu^2u_{xx}\quad\text{for}\quad x>w_t, t\in(0,T), \tag{1.6}$$

where $\nu^2=(2-\sigma^2)/\sigma^2$, which is not very different from (1.1).

Due to Theorems 1.1 and 3.2 it is natural to ask with which rate the boundary values are taken if the boundary data are smooth. We show some rather strong restrictions on the rate in section 5, the results of which are based on Theorem 1.2 and the properties of an explicit barrier function introduced in section 4. Actually again we need not Theorem 1.2 but only the fact that the left-hand side of (1.4) is strictly negative. In section 5 we show that the modulus of continuity of solutions to (1.1) even with smooth g can be as bad as we wish if ν is sufficiently small, and the same happens to (1.5) if σ^2 is below but sufficiently close to 2. In this way we get natural restrictions showing what cannot be proved.

The properties of the barrier function from section 4 also allow us to give a proof of the following deterministic result saying that the Hölder boundary regularity of

solutions to $u_t + (1/2)\nu^2 u_{xx} = 0$ for a ν implies the regularity for all ν. By C we denote the set of all real-valued continuous functions $x.$ given on $[0,\infty)$.

THEOREM 1.3. *Let a deterministic function $w.$ belong to C, let $\nu, \nu_0 \in (0,\infty)$, and let v be a Perron's or probabilistic solution of*

$$v_t(t,x) + \frac{1}{2}\nu_0^2 v_{xx}(t,x) = 0, \quad (t,x) \in Q(w.),$$

$$v(t,w_t) = g_0(t,w_t) \quad \text{for} \quad t \in [0,T), \quad u(T,x) = g_0(T,x) \quad \text{for} \quad x \geq w_T,$$

where g_0 is a nonnegative bounded continuous function such that $g_0(T,x) > 0$ for an $x > w_T$. Assume that for a $\lambda \in (0,1]$ and all $t \in [0,T]$ and $x \in (0,1]$, we have

$$v(t, x + w_t) \leq x^\lambda. \tag{1.7}$$

Finally, let u be a Perron's or probabilistic solution of (1.1)–(1.2) *with a bounded continuous function g. Then* (1.3) *holds for any $t_0 \in [0,T]$.*

Notice that if instead of (1.7) we just assume that (1.3) holds with v in place of u for any $t_0 \in [0,T]$, then the conclusion of Theorem 1.3 becomes false. This is easily shown by referring to Khinchin's law of the iterated logarithm.

The proof of Theorem 1.3, which we present in section 6, is based on two components. The first one is just a standard fact (see Lemma 2.1) that the solution of (1.1)–(1.2) is continuous at a point (t_0, w_{t_0}) on the lateral boundary if there is a parabola $t \geq t_0 + a^2(x - w_{t_0})^2$ with the pole at this point such that the boundary has common points with the (interior of the) parabola in any small neighborhood of the pole. This fact is quite similar to the exterior cone condition for elliptic equations, and, actually, it also holds in the situation when ν^2 is any Borel bounded function of (t,x) bounded away from zero. Although the proof of this generalization follows the same lines and is not much longer than that in the case of constant ν^2, we chose not to give it for the sake of brevity and because of the particular applications of our results to SPDEs. The referee of this paper kindly communicated to us how the continuity of solutions on the boundary for constant ν^2 can also be derived from the general necessary and sufficient condition of regularity proved in [4].

The second component is provided by the following result.

THEOREM 1.4. *Under the assumptions of Theorem* 1.3 *there exists a constant $c_0 \in (0,\infty)$ depending only on ν_0 and λ (see Remark* 6.1*) such that for all $t \in [0,T)$,*

$$\varlimsup_{h \downarrow 0} \frac{w_{t+h} - w_t}{\sqrt{h}} \geq -c_0. \tag{1.8}$$

We prove this theorem in section 6. Observe that if the left-hand side of (1.8) were too big negative at a point $t_0 \in [0,T)$, then there would exist a parabola $t \geq t_0 + a^2(x - w_{t_0})^2$ with large a such that its sufficiently small piece near the pole was inside of Q. Then the barrier from section 4 would imply that an opposite inequality holds in (1.7) for small $x > 0$ with $\lambda > 0$ tending to zero as $a \to \infty$.

Also in section 6, general existence and embedding theorems from the theory of SPDEs and the above discussed relation between (1.5) and (1.6) allow us to get Hölder continuity for solutions of the latter equation, which after combining with Theorem 1.4 allows us to give a proof of a relaxed version of the law of square root (of course, without using this law). This proof does a poor job in what concerns specifying the constant c_0, but the fact that $c_0 < \infty$ is explained qualitatively without computations.

By the "relaxed version" we mean the following fact in which $w_\cdot$ is the one-dimensional Wiener process from the beginning of the article.

THEOREM 1.5. (i) *There exists a constant $c_0 < \infty$ such that for any constant $T \in (0,\infty)$, we have (a.s.)*

$$\sup_{t\in(0,T]} \varliminf_{h\downarrow 0} \frac{w_t - w_{t-h}}{\sqrt{h}} \le c_0. \tag{1.9}$$

(ii) *One can take $c_0 = 2\sqrt{\pi}$. (Theorem 1.2 says that one can take $c_0 = 1$ and replace $\le$ with $=$ in (1.9).)*

2. Proof of Theorem 1.1. We start with the standard result we alluded to before Theorem 1.4.

LEMMA 2.1. *Take a deterministic function $w_\cdot \in C$ and assume that for any $t \in [0,T)$,*

$$\varliminf_{h\downarrow 0} \frac{w_{t+h} - w_t}{\sqrt{h}} =: -c(t) > -\infty. \tag{2.1}$$

Let g be a bounded continuous function. Then the probabilistic solution u of (1.1)–(1.2) satisfies (1.1) and assumes the boundary data (1.2). In particular, for any $t_0 \in [0,T]$, (1.3) holds.

Proof. Let B_t be a one-dimensional Wiener process (remember that here $w_\cdot$ is a fixed element of C). For $t, x \in \mathbb{R}$ also define

$$\tau(t,x) = \inf\{s > 0 : (t+s, x + \nu B_s) \notin Q\},$$

$$u(t,x) = Eg(t + \tau(t,x), x + \nu B_{\tau(t,x)}),$$

so that u is the probabilistic solution of (1.1)–(1.2).

Since B_t is a strong Markov process, for any box $P := (a,b) \times (c,d) \subset Q$ and $(t,x) \in P$, we have

$$u(t,x) = Eu(t + \gamma(t,x), x + \nu B_{\gamma(t,x)}),$$

where $\gamma(t,x) = \inf\{s > 0 : (t+s, x + \nu B_s) \notin P\}$. Therefore, it follows from [5] that u is infinitely differentiable in P and satisfies (1.1) there. Since $P \subset Q$ is arbitrary, u satisfies (1.1) in Q.

Now the only issue is that of the boundary values. Of course, as $t \uparrow T$ we have $T - t \ge \tau(t,x) \to 0$ and $u(t,x) \to g(t,x)$ for $x \ge w_T$ due to the continuity of g. However, the issue of the lateral boundary values is more delicate.

Observe that obviously $\tau(t,x)$ is a decreasing function of x, in particular, $\tau(t,x) \ge \tau(t,w_t)$ for $x \ge w_t$, and

$$u(t,x) \to Eg(t + \tau(t,w_t+), w_t + \nu B_{\tau(t,w_t+)})$$

as $x \downarrow w_t$, where

$$\tau(t,w_t+) \ge \tau(t,w_t).$$

Notice that by Blumenthal's 0-1 law, for any (t,x), we have that $P(\tau(t,x) = 0)$ equals 0 or 1. Furthermore, obviously, for $t \in [0,T)$, for $x = w_t$, and for any $h \in (0, T-t)$, we have

$$P(x + \nu B_h \le w_{t+h}) \le P(\tau(t,x) \le h).$$

Owing to (2.1), we can choose $h \in (0, T-t)$ as small as we like so that $w_{t+h} \geq x - 2c(t)\sqrt{h}$, and then

$$P(\nu B_h \leq -2c(t)\sqrt{h}) = P(x + \nu B_h \leq x - 2c(t)\sqrt{h}) \\ \leq P(x + \nu B_h \leq w_{t+h}) \leq P(\tau(t,x) \leq h).$$

Here the first expression is independent of h and is strictly positive. Hence

$$P(\tau(t, w_t) = 0) = \lim_{h \downarrow 0} P(\tau(t, w_t) \leq h) > 0,$$

which implies that $P(\tau(t, w_t) = 0) = 1$ by Blumenthal's 0-1 law.

The proof of the lemma would have stopped here if we knew a reference showing that the well-known boundary regularity results (see, for instance, [3], [6]) were true not only for strong Feller processes but also for processes with strong Feller resolvent. The author could not find such a reference in the literature and this is why, to show that for any $t \in [0, T)$, not only $\tau(t, w_t) = 0$ (a.s.) but also $\tau(t, w_t+) = 0$ (a.s.), we just repeat a standard argument from the theory of Markov processes (see, for instance, [3]). Define

$$v(t,x) = E\tau(t,x)$$

and notice that, for $s > 0, t \in [0, T)$, and $t + s \leq T$, by Markov property

$$Ev(t + s, x + \nu B_s) = E\tau_s(t,x) - s,$$

where

$$\tau_s(t,x) = \inf\{r > s : (t + r, x + \nu B_r) \notin Q\}.$$

Hence

$$E\tau_s(t,x) = s + \frac{1}{\sqrt{2\pi\nu^2 s}} \int_{\mathbb{R}} v(t+s, y) e^{-(y-x)^2/(2\nu^2 s)}\, dy,$$

which shows that $E\tau_s(t,x)$ is continuous in x. Furthermore, obviously $\tau_s(t,x) \downarrow \tau(t,x)$ as $s \downarrow 0$, and by the dominated convergence theorem (remember that $\tau_s(t,x) \leq T - t$) we have $E\tau_s(t,x) \downarrow E\tau(t,x)$. Since $E\tau_s(t,x)$ is continuous in x, we conclude that $E\tau(t,x)$ is upper semicontinuous in x for any $t \in [0, T)$. In particular,

$$E\tau(t, w_t+) = \overline{\lim_{x \downarrow w_t}} E\tau(t,x) \leq E\tau(t, w_t).$$

Here the right-hand side is zero by the above. Thus,

$$v(t, w_t+) = E\tau(t, w_t+) = 0$$

indeed. We assumed that $t < T$, but since $\tau(T, x) \equiv 0$, our conclusion holds for all $t \in [0, T]$

Now, observe that by Itô's formula, for $h(t,x) = t$, we have

$$h(t,x) = -v(t,x) + Eh(t + \tau(t,x), x + \nu B_{\tau(t,x)}),$$

so that the argument in the beginning of the proof shows that $v(t,x)$ is a continuous function in $Q \cup \{(0,x) : x > 0\}$. It is also continuous in $Q \cup \{(T,x) : x \geq w_T\}$

because $v(t,x) \le T - t \to 0$ as $t \uparrow T$. Finally, the functions $v(t, x + w_t)$ are continuous in $t \in [0,T]$ if $x > 0$ and for each t decrease to zero as $x \downarrow 0$. By Dini's theorem $v(t, x + w_t) \to 0$ as $x \downarrow 0$ uniformly in $t \in [0,T]$. In particular, $\tau(t,x) \to 0$ in probability as $Q \ni (t,x) \to (t_0, w_{t_0})$. After that, (1.3) follows from the definition of u and the continuity of g. The lemma is proved.

Remark 2.2. One can show that for each particular $t_0 \in [0,T)$, (1.3) holds if (2.1) holds only for $t = t_0$.

Now, Theorem 1.2 says that the conditions of Lemma 2.1 hold with $c(t) = 1$ for almost any trajectory $w.$ of the Wiener process. Hence, for almost any trajectory $w.$, the conclusion of this lemma holds no matter which continuous and bounded g we take. This is exactly the assertion of Theorem 1.1.

In the same way we get Theorem 1.1 from Theorem 1.5 after noticing that since $w_T - w_{T-t}$ is a Wiener process on $[0,T]$, (1.9) is equivalent to saying that (a.s.)

$$\inf_{t \in [0,T)} \varlimsup_{h \downarrow 0} \frac{w_{t+h} - w_t}{\sqrt{h}} \ge -c_0.$$

3. An application to SPDEs. Let $\sigma \ge 0$ be a number such that $\sigma^2 < 2$, and let ζ be a nonnegative function satisfying $\zeta \in C_0^\infty(0,\infty)$ and $\zeta(x) = 1$ for $x \in [1,2]$. Consider equation (1.5) for $t \in (0,T)$ and $x > 0$ with zero lateral condition and with $f(0,x) = \zeta(x)$ for $x > 0$.

Of course, by solution f we mean an appropriately measurable and integrable function $f = f(\omega, t, x)$ such that for any test function $\psi \in C_0^\infty(0,\infty)$,

$$(f(t,\cdot),\psi) = (\zeta,\psi) + \int_0^t (f(s,\cdot),\psi_{xx})\,ds - \sigma \int_0^t (f(s,\cdot),\psi_x)\,dw_s \tag{3.1}$$

(a.s.) for all $t \in [0,T]$, and where $(\cdot,\cdot)$ is the scalar product in L_2. This shows how the equation and the initial condition are understood. The condition that $f = 0$ on $x = 0$ is reflected in the requirement that f belong to an appropriate Banach space. To be a little bit more specific, if $r \ge 2$ and $\theta \in \mathbb{R}$ satisfy

$$1 - \frac{2}{(r-1)\sigma^2 + 2} < \frac{\theta}{r} < 1, \tag{3.2}$$

then by Theorem 3.3 of [11] equation (1.5) with given initial data admits a (unique) solution belonging to the class $\mathfrak{H}^\gamma_{r,\theta}(T)$ for all $\gamma \in \mathbb{R}$. Here γ is the number of derivatives of f in x, r is the power of summability in (ω,t,x), and θ is "responsible" for the rate with which the derivatives of f in x are allowed to blow up near $x = 0$. The precise definition of $\mathfrak{H}^\gamma_{r,\theta}(T)$, which we do not need in the present article, is given in [10] and [11]. The following fact is a direct consequence of the results in [12], [7], and [9].

LEMMA 3.1. *For almost any ω, the function $f(t,x)$ is infinitely differentiable in x, any of its derivatives with respect to x are continuous in (t,x) in the region $t \in [0,T], x > 0$, and, for any $x > 0$, (1.5) holds for $t \in (0,T)$.*

It may be worth noting that, of course, the assertions of the lemma refer to an appropriate modification of the solution rather than to the solution itself. Here is the main result of this section also stated for the modification.

THEOREM 3.2. *For almost any ω, we have*

$$\sup_{t \in [0,T]} |f(t,x)| \to 0 \quad \text{as} \quad x \downarrow 0.$$

What follows below in this section is aimed at proving this theorem. On the space C with Wiener measure W, introduce the coordinate process $x_t(x.) := x_t$, which is a Wiener process. For $t \geq 0$, $x \in \mathbb{R}$, and $x., y. \in C$, define

$$\tau(t, x, x., y.) = t \wedge \inf\{s \geq 0 : x + \nu x_s \leq y_{t-s}\},$$

where $y_r := 0$ for $r \leq 0$ and $\nu = \sigma^{-1}\sqrt{2 - \sigma^2}$. Then the probabilistic solution of (1.6) with zero lateral condition and with initial condition $\eta(x) := \zeta(\sigma x)$ is given by

$$u(t, x) = u(\omega, t, x) := v(w.(\omega), t, x),$$

where

$$v(y., t, x) := \int_C I_{\tau(t,x,x.,y.)=t}\, \eta(x + \nu x_t)\, W(dx.).$$

From Theorem 1.1, upon noticing that $u(T - t, x + w_T)$ solves (1.1)–(1.2) with $w_{T-t} - w_T$ in place of w_t and with $g = 0$ on the lateral boundary, we immediately get that $u(t, x/\sigma + w_t) \to 0$ as $x \downarrow 0$ uniformly in $t \in [0, T]$. Therefore, to prove the present theorem it suffices to show that $f = \bar{f}$ (a.s.), where

$$\bar{f}(t, x) := u(t, x/\sigma + w_t).$$

Observe that $\tau(t, x, x., y.)$ is a lower semicontinuous function of $\nu x. - y_{t-.}$ for each (t, x). Therefore by Fubini's theorem $v(y., t, x)$ is a Borel function of $y.$ and $u(t, x)$ is a random variable. Furthermore, $v(y., t, x)$ will not change if we change y_r for $r > t$. Therefore, $v(y., t, x)$ is $\sigma(y_r : r \leq t)$-measurable and $u(t, x)$ is $\mathcal{F}_t^w$-adapted, where $\mathcal{F}_t^w = \sigma(w_r : r \leq t)$. In addition u satisfies (1.6) and hence is infinitely differentiable in $[0, T] \times (0, \infty)$ and, in particular, continuous in (t, x) for any ω. Hence $\bar{f}$ satisfies the measurability properties required for solutions of (1.5). By using (1.6) and the fact that u is infinitely differentiable in (t, x) and by using the Itô–Wentzell formula, we get that with probability one, for all $t \in [0, T]$ and $x > 0$,

$$\bar{f}(t, x) = \zeta(x) + \int_0^t \bar{f}_{xx}(s, x)\, ds + \sigma \int_0^t \bar{f}_x(s, x)\, dw_s. \tag{3.3}$$

Next we want to pass from this pointwise equation to (3.1). For $\varepsilon > 0$ denote by $\Gamma(\varepsilon, w.)$ the two-dimensional ε-neighborhood of the graph of $w_t, t \in [0, T]$. Obviously $0 \leq u \leq 1$ and for such solutions of the heat equation it is known that, given $\varepsilon > 0$, any derivative of u with respect to (t, x) is bounded in

$$(t, x) \notin \Gamma(\varepsilon, w.), \quad x > w_t, \quad t \in [0, T],$$

by a constant depending only on ε and the order of the derivative. Therefore, if we take a $\psi \in C_0^\infty(0, \infty)$, then with probability one any derivative in x of $\bar{f}(t, x)$ is bounded and continuous in (t, x) whenever $x \in \operatorname{supp} \psi$ and $t \in [0, T]$. This allows us to use the stochastic Fubini's theorem and obtain (3.1) for $\bar{f}$ after multiplying (3.3) by $\psi(x)$ and integrating with respect to x.

Since f also satisfies (3.1) and there is uniqueness, to show that $f = \bar{f}$ we prove that $\bar{f}$ belongs to $\mathfrak{H}_{r,\theta}^\gamma(T)$.

LEMMA 3.3. *For $r \geq 1$ and $\theta \in (r - 1, r)$, we have $\bar{f} \in \mathbb{L}_{r,\theta-r}(T)$, that is,*

$$E \int_0^T \int_0^\infty x^{\theta-r-1} |\bar{f}(t, x)|^r\, dx dt < \infty.$$

Proof. Fix $t \in (0,T)$ and $x > 0$. By Hölder's inequality

$$E|\bar{f}(t,x)|^r \leq \int_\Omega \int_C I_{\gamma(t,x)=t}\, \zeta^r(x + \sigma\nu x_t + \sigma w_t)\, W(dx.)P(d\omega), \tag{3.4}$$

where

$$\begin{aligned}\gamma(t,x) &= \tau(t, x/\sigma + w_t, x., w.)\\ &= t \wedge \inf\{s \geq 0 : x/\sigma + w_t + \nu x_s \leq w_{t-s}\}\\ &= t \wedge \inf\{s \geq 0 : x + \sigma m_t + \sqrt{2-\sigma^2}\, x_s \leq 0\},\end{aligned}$$

and $m_s := w_t - w_{t-s}$ is a Wiener process with respect to $s \in [0,t]$. Denote

$$\sqrt{2}B_s = \sigma m_t + \sqrt{2-\sigma^2}\, x_s$$

and notice that B_t is a Wiener process on $\Omega \times C$,

$$\gamma(t,x) = t \wedge \inf\{s \geq 0 : x + \sqrt{2}B_s \leq 0\}.$$

Now, in a common abuse of notation we write E for the expectation sign on $\Omega \times C$ and rewrite (3.4) as

$$E|\bar{f}(t,x)|^r \leq E I_{\gamma(t,x)=t}\zeta^r(x + \sqrt{2}\, B_t) =: h(t,x).$$

We recognize h as the probabilistic solution of the heat equation $h_t = h_{xx}$, $t \in (0,T), x > 0$, with zero boundary condition and with initial condition ζ^r. From well-known estimates for solutions of such problems, we get that

$$|h(t,x)| \leq Nx \quad \text{for} \quad x \leq 1, \qquad |h(t,x)| \leq Ne^{-x} \quad \text{for} \quad x \geq 1,$$

where the constant N is independent of $t \in [0,T], x > 0$. The reader preferring probabilistic proofs can get the same estimates after noticing that the event $\gamma(t,x) = t$ coincides (a.s.) with $\inf_{s \leq t} B_s > -x/\sqrt{2}$ and the distribution of $(B_t, \inf_{s\leq t} B_s)$ is well known.

The above estimate of $E|\bar{f}(t,x)|^r$ immediately implies the assertion of the lemma. The lemma is proved.

From [10], [11] we know what the differentiating does to functions from $\mathbb{H}^\gamma_{r,\theta}(T)$ and obtain the following.

COROLLARY 3.4. *We have $\bar{f}_x \in \mathbb{H}^{-1}_{r,\theta}(T)$, $\bar{f}_{xx} \in \mathbb{H}^{-2}_{r,\theta+r}(T)$, so that $\bar{f}$ is an $\mathfrak{H}^0_{r,\theta}(T)$-solution of (1.5) if $\theta \in (r-1, r)$ and $r \geq 2$.*

Now since the initial data is smooth, the results of [10] show that $\bar{f} \in \mathfrak{H}^\gamma_{r,\theta}(T)$ for any γ if $\theta \in (r-1,r)$ and $r \geq 2$. Finally, uniqueness theorems (see, for instance, Lemma 4.3 of [10]) prove that no matter to which space $\mathfrak{H}^\gamma_{p,\tau}(T)$ the function f belongs, we have $f = \bar{f}$ for almost all $(\omega,t,x) \in \Omega \times [0,T] \times (0,\infty)$. Since both functions are continuous in (t,x), we conclude that (a.s.) $f(t,x) = \bar{f}(t,x)$ for all $t \in [0,T], x > 0$, and this brings the proof of Theorem 3.2 to an end.

4. A barrier function. For $c \geq 0$ consider the domain

$$D_c = \{(t,x) : t \in (0,1), x > -c\sqrt{1-t}\}$$

with the lateral boundary being the parabola

$$\Gamma_c := \{(t,x) : t \in [0,1], x = -c\sqrt{1-t}\}.$$

We will be interested in finding a nonnegative nontrivial solution of the heat equation

$$u_t = \frac{1}{2}u_{xx} \quad \text{in} \quad D_c \tag{4.1}$$

which is continuous in $\bar{D}_c$ and vanishes on Γ_c. If we look for u in the form $u(t,x) = f(t)\phi(y)$, where $y = x/\sqrt{1-t}$, then, by noticing that

$$u_t = f'\phi + \frac{y}{2(1-t)}f\phi', \quad u_x = \frac{1}{\sqrt{1-t}}f\phi', \quad u_{xx} = \frac{1}{1-t}f\phi'',$$

we find that u satisfies (4.1) if

$$f'\phi = \frac{f}{2(1-t)}\,(\phi'' - y\phi'),$$

that is, if there is a number λ such that

$$\frac{f'}{f} = -\frac{\lambda}{2(1-t)}, \quad t \in (0,1), \qquad \phi'' - y\phi' = -\lambda\phi, \quad y > -c.$$

The following argument is based on the results of [14]. According to [14] the function

$$\psi_0(\lambda,x) := \int_0^\infty p(x,r)r^{-\lambda-1}\,dr, \qquad p(x,r) := \exp(-rx - r^2/2),$$

satisfies $\phi'' - y\phi' = -\lambda\phi$ for all $x \in \mathbb{R}$ if $\lambda < 0$. By repeatedly integrating by parts, we see that for those λ and any $n = 0,1,2,\ldots$, we have $\psi_0(\lambda,x) = N(\lambda)\psi_n(\lambda,x)$, where

$$\psi_n(\lambda,x) := \int_0^\infty r^{n-\lambda-1}\frac{\partial^n}{(\partial r)^n}\,p(x,r)\,dr,$$

and $N(\lambda) \neq 0$. Hence $\psi_n(\lambda,x)$ also satisfies $\psi_n'' - y\psi_n' = -\lambda\psi_n$ for all $x \in \mathbb{R}$ if $\lambda < 0$. However, since both parts of the equation are obviously analytic in λ if $\Re\lambda < n$, the equation holds whenever $\lambda < n$. (Actually, this can be checked out by directly integrating back by parts and noticing that by Leibniz's formula

$$p_r^{(n+1)}(x,r) = -((x+r)p(r,x))_r^{(n-1)} = -(x+r)p_r^{(n)}(x,r) - np_r^{(n-1)}(x,r).\ )$$

We will be interested in $n = 1$ and $0 < \lambda < 1$ when

$$\psi_1(\lambda,x) = \int_0^\infty r^{-\lambda}\frac{\partial}{\partial r}\,[p(x,r) - 1]\,dr = -\lambda\phi(\lambda,x),$$

where

$$\phi(\lambda,x) := \int_0^\infty [1 - e^{-xr-r^2/2}]r^{-\lambda-1}\,dr. \tag{4.2}$$

It is seen that $\phi(\lambda,x)$ is a strictly increasing function of x, $\phi(\lambda,x) \geq \phi(\lambda,0) > 0$ for $x \geq 0$, $\phi(\lambda,x) \to \pm\infty$ as $x \to \pm\infty$. Hence, for any $\lambda \in (0,1)$, there exists a unique point $c(\lambda) > 0$ such that $\phi(\lambda,-c(\lambda)) = 0$. In addition,

$$\phi_x(\lambda,x) = \int_0^\infty e^{-xr-r^2/2}r^{-\lambda}\,dr > 0, \tag{4.3}$$

and, for $c = c(\lambda)$,

$$\phi_\lambda(\lambda, -c) = -\int_0^\infty [1 - e^{cr - r^2/2}] r^{-\lambda - 1} \ln(r/2c)\, dr < 0.$$

(Notice that the last integrand is nonnegative. The author learned this observation from M. Safonov.) It follows that $c(\lambda)$ is continuously differentiable and strictly decreasing for $\lambda \in (0,1)$. Therefore, the function $\lambda \to c(\lambda)$ is invertible. In this way we obtain part of the results in [14] where the properties of Sturm–Liouville problems are used. The remaining part of the following result is taken directly from [14]. At this point it is also worth noting that similar results for equation $\phi'' - x\phi' = -\lambda\phi$ not on half lines but on finite intervals are given in [13].

LEMMA 4.1. *For any $c > 0$ there exists a unique $\lambda = \lambda(c) \in (0,1)$ such that*

$$\int_0^\infty [1 - e^{cr - r^2/2}] r^{-\lambda - 1}\, dr = 0. \tag{4.4}$$

The function $\lambda(c)$ is differentiable and strictly decreasing on $(0,\infty)$. Finally ($a \sim b$ means $a/b \to 1$), we have $1 - \lambda(c) \sim c\sqrt{2/\pi}$ as $c \downarrow 0$ and $\lambda(c) \sim (2\pi)^{-1/2} ce^{-c^2/2}$ as $c \to \infty$.

Remark 4.2. There is a very indirect argument showing that (4.4) for $\lambda \in (0,1)$, $c > 0$ is equivalent to the equation

$$\int_0^1 \left[\frac{1}{\sqrt{1 - r^2}} e^{c^2 r/(1+r)} - 1 \right] r^{-\lambda - 1}\, dr = \frac{1}{\lambda}\,.$$

The author does not know any elementary proof of the equivalence.

Now we are ready to introduce a barrier function.

LEMMA 4.3. *Take $c \in (0,\infty)$, define $\lambda \in (0,1)$ as the unique solution of (4.4), and let*

$$v(t,x) := \int_0^\infty [1 - e^{-rx - (1-t)r^2/2}] r^{-\lambda - 1}\, dr, \quad (t,x) \in \bar{D}_c.$$

Then

(i) *v is infinitely differentiable in D_c, $v_x > 0$, $v_{xx} < 0$, and $v_t < 0$ in D_c, and v satisfies (4.1);*

(ii) *v is continuous in $\bar{D}_c$, increases and is concave in x, decreases in t, $v > 0$ in D_c, $v = 0$ on Γ_c;*

(iii) *$v(1,x) = Nx^\lambda$ for $x \geq 0$, where $N = \lambda^{-1}\Gamma(1-\lambda)$, and $v(t,0) = N_1(1-t)^{\lambda/2}$ for $t \in [0,1]$, where the constant $N_1 \in (0,\infty)$.*

Proof. Take the function $\phi(\lambda, x)$ according to (4.2). Then the substitution $r \to r\sqrt{1-t}$ shows that $(1-t)^{\lambda/2}\phi(\lambda, x/\sqrt{1-t}) = v(t,x)$ if $(t,x) \in \bar{D}_c$ and $t \neq 1$. This, together with what has been said before in this section, immediately implies all assertions in (i), perhaps apart from the ones concerning the signs of derivatives. However, the signs of (all) derivatives in x are obtained from (4.3), and then we get $v_t = (1/2)v_{xx} < 0$.

We also see that to prove (ii) we need only prove that v is continuous at the top of D_c. The continuity at points $t = 1, x > 0$ follows from the dominated convergence theorem. To consider the point $(1,0)$, observe that for any $\alpha > 0$,

$$\int_0^\infty [1 - e^{-\alpha r}] r^{-\lambda - 1}\, dr = \alpha^\lambda \int_0^\infty [1 - e^{-r}] r^{-\lambda - 1}\, dr = N\alpha^\lambda. \tag{4.5}$$

Then, for $(t,x) \in \bar{D}_c$, $t \neq 1$, we find

$$\begin{aligned} v(t,x) - v(1,0) = v(t,x) &= v(t,x) - \int_0^\infty [1 - e^{cr\sqrt{1-t} - r^2(1-t)/2}] r^{-\lambda-1}\, dr \\ &= \int_0^\infty e^{cr\sqrt{1-t} - r^2(1-t)/2} [1 - e^{-r(x+c\sqrt{1-t})}] r^{-\lambda-1}\, dr \\ &\le e^{c^2/2} \int_0^\infty [1 - e^{-r(x+c\sqrt{1-t})}] r^{-\lambda-1}\, dr = N(x + c\sqrt{1-t})^\lambda \to 0 \end{aligned}$$

as $(t,x) \to (1,0)$. This finishes the proof of (ii).

Obviously, assertion (iii) follows immediately from (4.5). The lemma is proved.

Now comes the main result of this section. It says that the modulus of continuity of solutions to the heat equation in a domain bounded by half parabola *is* affected by the slope of the parabola. Of course, we consider parabolas only with axes parallel to the t-axis and directed down with respect to the t-axis. This, together with Lemma 4.1, shows that if we have a nonnegative solution which is, say, λ-Hölder continuous in a domain, then the domain cannot contain parabolas that are too "wide" with poles on the boundary and the critical "width" is determined by λ.

LEMMA 4.4. *Let $c \in (0,\infty)$, $t_0 \in (0,\infty)$, and $x_0 \in \mathbb{R}$. Take an $a \in (0, t_0]$ and denote*

$$G_{c,a}(t_0,x_0) = \{(t,x) : t_0 - a < t < t_0, -c\sqrt{(t_0 - t)} < x - x_0 < 2\sqrt{a}\}.$$

Let $u(t,x)$ be a bounded continuous function given in $\bar{G}_{c,a}(t_0,x_0)$ and satisfying

$$\frac{1}{2} u_{xx} - u_t \le 0 \tag{4.6}$$

in $G_{c,a}(t_0,x_0)$ in the classical sense. Assume that $u \ge 0$ and $u \not\equiv 0$ in $G_{c,a}(t_0,x_0)$. Then there exists a constant $\delta > 0$ such that

$$u(t_0,x) \ge \delta (x - x_0)^{\lambda(c)}$$

for $0 \le x - x_0 \le \sqrt{a}$, where $\lambda(c)$ is introduced in Lemma 4.1.

Proof. Notice that the function

$$u(a(t-1) + t_0, x\sqrt{a} + x_0)$$

satisfies the assumption of the lemma with $t_0 = a = 1$ and $x_0 = 0$. Furthermore, the assertion of the lemma is also easily rewritten in terms of this new function. Therefore, without losing generality we assume that $t_0 = a = 1$ and $x_0 = 0$, so that u satisfies (4.6) in

$$G_c := G_{c,1}(1,0).$$

By the Harnack inequality, we have $u(t,1) > 0$ in a closed left neighborhood of 1. Then simple barriers show that for any $s_0 \in (0,1)$ sufficiently close to 1, there exists an $\varepsilon > 0$ such that

$$u(s_0,x) \ge \varepsilon (x + c\sqrt{(1 - s_0)})$$

for all $x \in [-c\sqrt{(1-s_0)}, 1]$ and $u(t,1) \ge \varepsilon$ for $t \in [s_0, 1]$. We fix appropriate $s_0 \in [3/4, 1)$ and $\varepsilon > 0$ and take $\lambda = \lambda(c) \in (0,1)$.

We also take the functions v from Lemma 4.3 and observe that $v(s_0,x)$ is a smooth function vanishing at $x=-c\sqrt{(1-s_0)}$. Therefore, there is a constant $\gamma>0$ such that

$$\gamma v(s_0,x)\leq\varepsilon(x+c\sqrt{(1-s_0)})\leq u(s_0,x).$$

By reducing $\gamma>0$ if necessary we can achieve the inequality $\gamma v(t,1)\leq u(t,1)$ for all $t\in[s_0,1]$. Then the inequality $v\leq u$ holds on the parabolic boundary of G_c, and by virtue of (4.6) and the maximum principle we have $\gamma v\leq u$ everywhere in $\bar{G}_c$. In particular, $u(1,x)\geq\gamma v(1,x)$ for $0\leq x\leq 1$, which, owing to Lemma 4.3(iii), yields our assertion. The lemma is proved.

Remark 4.5. One can get estimates for $u(t_0,x)$ from above as well. Let $c\in(0,\infty)$, $t_0\in(0,\infty)$, $a\in(0,t_0]$, and $x_0\in\mathbb{R}$. Let $u(t,x)$ be a bounded continuous function given in $\bar{G}_{c,a}(t_0,x_0)$ and satisfying

$$\frac{1}{2}u_{xx}-u_t\geq 0$$

in $G_{c,a}(t_0,x_0)$ in the classical sense. Also assume that $u\leq 0$ for $x-x_0=-c\sqrt{(t_0-t)}$ if $t_0-a\leq t\leq t_0$. Then by using the maximum principle, one easily obtains that there exists a constant $K>0$ such that $u(a(t-1)+t_0,x\sqrt{a}+x_0)\leq Kv(t,x)$ in the intersection of $\bar{G}_c$ with a neighborhood of $(1,0)$. It follows that there exists a constant N such that $u(t_0,x)\leq N(x-x_0)^{\lambda(c)}$ for $0\leq x-x_0\leq\sqrt{a}$, where $\lambda(c)$ is introduced in Lemma 4.1.

5. Lower estimates on the modulus of continuity of u and f on the boundary. In view of Lemma 4.1 the following theorem shows that the Hölder exponent of solutions to (1.1)–(1.2) can be extremely small if ν is small.

THEOREM 5.1. *Let $c>0$ be a constant, and let $0<\nu c<1$. Then there exists a measurable set $\Omega'\subset\Omega$ such that $P(\Omega')=1$ and for each $\omega\in\Omega'$, there exists an everywhere dense subset S of $[0,T]$ such that for any $t_0\in S$ and nonnegative continuous bounded g, satisfying $g(T,x)\not\equiv 0$ for $x>w_T$ (and, say, equal to zero whenever $x=w_t$ and $t\in[0,T]$), we have*

$$\lim_{x\downarrow w_{t_0}}\frac{u(t_0,x)}{(x-w_{t_0})^{\lambda(c)}}=\infty, \tag{5.1}$$

where u is the probabilistic solution of problem (1.1)–(1.2).

Proof. Obviously, it suffices to show that (a.s.) on each dyadic subinterval $[Tk2^{-n},T(k+1)2^{-n}]$ of $[0,T]$ there is a point t_0 such that (5.1) holds for any solution of (1.1) on $[Tk^{-n},T(k+1)2^{-n})$ in place of $[0,T)$ satisfying $u(T(k+1)2^{-n},x)\geq 0$ and $u(T(k+1)2^{-n},x)\not\equiv 0$. Due to self similarity of the heat equation and the Wiener process, the problem for each subinterval reduces to the one for $[0,1]$, and since there are only countably many dyadic subintervals, it suffices to prove the existence of S which is not everywhere dense but rather just nonempty.

Notice that

$$I:=\lim_{n\to\infty}\inf_{t\in[0,T/2]}\sup_{h\in(0,1/n]}\frac{w_{t+h}-w_t}{\sqrt{h}}=\inf_{n\geq 1}\inf_{t\in[0,T/2]}\sup_{h\in(0,1/n]}\frac{w_{t+h}-w_t}{\sqrt{h}}$$

$$=\inf_{t\in[0,T/2]}\inf_{n\geq 1}\sup_{h\in(0,1/n]}\frac{w_{t+h}-w_t}{\sqrt{h}}=\inf_{t\in[0,T/2]}\overline{\lim_{h\downarrow 0}}\frac{w_{t+h}-w_t}{\sqrt{h}}:=J.$$

Owing to (1.4), we have $J=-1$ (a.s.). Hence, there is a set Ω' of full probability on which $I=-1$. We take any $\omega\in\Omega'$ and a $c'\in(c,\nu^{-1})$. Then there exists $n\geq 1$ and $t_0\in[0,T/2]$ such that for $w_\cdot=w_\cdot(\omega)$,

$$\sup_{h\in(0,1/n]}\frac{w_{t_0+h}-w_{t_0}}{\sqrt{h}}\leq -c'\nu$$

or, equivalently, $w_{t_0+h}\leq x_0-c'\nu\sqrt{h}$ for $h\in(0,1/n)$, where $x_0=w_{t_0}$. It follows that the function $v(t,x)=u(t/\nu^2,x)$ satisfies the equation $v_t+(1/2)v_{xx}=0$ for

$$x>x_0-c'\sqrt{t-t_0}$$

and $0<t-t_0<a$, where $a=\nu^2/n$. After changing variables $t\to T-t$ we transform the equation $v_t+(1/2)v_{xx}=0$ into $v_t=(1/2)v_{xx}$ and get the possibility to apply Lemma 4.4, which leads to $u(t_0,x)\geq\delta(x-x_0)^{\lambda(c')}$ for small $x-x_0>0$ and to

$$\lim_{x\downarrow x_0}\frac{u(t_0,x)}{(x-x_0)^{\lambda(c)}}=\lim_{x\downarrow x_0}\frac{u(t_0,x)}{(x-x_0)^{\lambda(c')}}\frac{1}{(x-x_0)^{\lambda(c)-\lambda(c')}}=\infty,$$

where the last conclusion follows from the inequality $\lambda(c)>\lambda(c')$, which holds because λ is a strictly decreasing function. The theorem is proved.

Remark 5.2. The results of [13] show that actually (a.s.) S has a nonzero Hausdorff dimension, which is independent of ω.

Our last result shows that there are some nontrivial restrictions on the modulus of continuity on the boundary of solutions of SPDEs.

THEOREM 5.3. *Let f be the function introduced in section* 3 *as a solution of* (1.5)*, and let a constant c satisfy*

$$0<c\sqrt{2-\sigma^2}<\sigma.$$

Then with probability one there exists a dense subset $S\in[0,T]$, which is unrelated with f and is such that for any $t_0\in S$, we have

$$\lim_{x\downarrow 0}\frac{f(t_0,x)}{x^{\lambda(c)}}=\infty.$$

Proof. Notice that as we have seen in the proof of Theorem 3.2 the function $u(t,x):=f(t,\sigma(x-w_t))$ satisfies $u_t=(1/2)\nu^2u_{xx}$ for $x>w_t$, $t\in[0,T]$, where $\nu^2=(2-\sigma^2)/\sigma^2$. After that it only remains to either reverse time and refer to Theorem 5.1 or just repeat the proof of this theorem again using Lemma 4.4 and avoid any time change, since the lemma is stated for the "usual" heat equation. The theorem is proved.

6. Proofs of Theorems 1.3, 1.4, and 1.5.

Proofs of Theorems 1.3 *and* 1.4. First we deal with Theorem 1.4. Take a $t_0\in[0,T)$ and observe that the function $u(t,x)=v(T-t,\nu_0x)$ satisfies $u_t=(1/2)u_{xx}$ in

$$R:=\{(t,x):t\in(0,T),x>\nu_0^{-1}w_{T-t}\}.$$

Then take any $c\in(0,\infty)$ such that

$$\lambda(c)<\lambda. \tag{6.1}$$

We notice that the point $(T-t_0,\nu_0^{-1}w_{t_0})$ is on the parabolic boundary of R and claim that no matter how small $a>0$ is,

$$G_{c,a}(T-t_0,\nu_0^{-1}w_{t_0})\notin R.$$

Indeed, otherwise there would exist $a>0$ such that for any $c'\in(0,c)$, the function u would be continuous in $\bar{G}_{c',a}(T-t_0,\nu_0^{-1}w_{t_0})$ and satisfy $u_t=(1/2)u_{xx}$ in $G_{c',a}(T-t_0,\nu_0^{-1}w_{t_0})$. Then by Lemma 4.4 we would have

$$0<\varliminf_{x\downarrow0}\frac{u(T-t_0,x+\nu_0^{-1}w_{t_0})}{x^{\lambda(c')}}=\varliminf_{x\downarrow0}\frac{v(t_0,\nu x+w_{t_0})}{x^{\lambda(c')}}=:I.$$

However, if $\lambda(c')<\lambda$, then $I=0$ due to condition (1.7). We get a contradiction since $\lambda(c)<\lambda$ and, owing to the continuity of $\lambda(c)$, one can indeed choose $c'\in(0,c)$ so that $\lambda(c')<\lambda$.

Our claim just proved is equivalent to saying that for any $a\in(0,T-t_0)$, there exists $h\in(0,a)$ such that

$$\nu_0^{-1}w_{t_0}-c\sqrt{h}\le\nu_0^{-1}w_{t_0+h}.$$

It follows that

$$\overline{\varlimsup_{h\downarrow0}}\ \frac{w_{t_0+h}-w_{t_0}}{\sqrt{h}}\ge-c\nu_0. \tag{6.2}$$

This finishes the proof of Theorem 1.4.

Theorem 1.3 follows immediately from Theorem 1.4 and Lemma 2.1.

Remark 6.1. Equation (6.2) holds whenever $c>0$ and condition (6.1) is satisfied. The latter can be rewritten as $c>c(\lambda)$, where we define $c(1)=c(1-)=0$ and, for other $\lambda\in(0,1)$, by $c(\lambda)$ we mean the function introduced in section 4 before Lemma 4.1.

It follows that (1.8) holds with $c_0=\nu_0c(\lambda)$. By the way, Lemma 4.3 shows that this value of c_0 is sharp as long as all possible boundaries are allowed.

Proof of Theorem 1.5. We will be using some properties of the solution f to (1.5) introduced in section 3 for the parameters $r\ge2$ and θ satisfying (3.2). Again the exact definition of the spaces $\mathfrak{H}^\gamma_{r,\theta}(T)$ is not at issue here. What is important for us is that by Theorem 4.1 of [9] or by Theorem 2.7 of [12], if $2/r<\eta\le1$ and $\gamma\in\mathbb{R}$, then

$$E\sup_{t\le T}||M^{\eta-1}f(t,\cdot)||^r_{H^\gamma_{r,\theta}}<\infty$$

and, finally, by Lemma 2.2 of [9] or by Theorem 3.1 of [7] that

$$E\sup_{t\le T}\sup_{x>0}|x^{\eta-1+\theta/r}f(t,x)|^r<\infty.$$

Due to the arbitrariness in the choice of η and θ, we get that

$$E\sup_{t\le T,x>0}(x^{\varepsilon-\beta}|f(t,x)|)^r<\infty,\qquad\beta=\frac{2}{(r-1)\sigma^2+2}-\frac{2}{r},$$

where $\varepsilon>0$ is as small as we like. Important at this moment is that there are $\sigma\in(0,1)$, $r>2$, and $\varepsilon\in(0,\beta)$ such that $\beta>0$ and the event

$$B(\beta-\varepsilon):=\left\{\omega:\overline{\varlimsup_{x\downarrow0}}\sup_{t\in[0,T]}\frac{|f(t,x)|}{x^{\beta-\varepsilon}}<\infty\right\}$$

has full probability. We fix some $\sigma \in (0,1)$, $r > 2$, and $\varepsilon \in (0, \beta)$ such that $\beta > 0$. By the way, if $\sigma \geq 1$, then $\beta < 0$ for any $r > 0$.

Next, we remember that according to section 3, for $\omega \in \Omega'$ with $P(\Omega') = 1$, the function $u(t,x) := f(t, \sigma(x - w_t))$ satisfies $u_t = (1/2)\nu^2 u_{xx}$ for $x > w_t$, $t \in (0,T)$, where $\nu = \sigma^{-1}\sqrt{2-\sigma^2}$. Hence, for $\omega \in \Omega'$, the function $v(t,x) := u(T-t,x)$ satisfies $v_t + (1/2)\nu^2 v_{xx} = 0$ for $x > w_{T-t}$, $t \in (0,T)$.

Furthermore, if $\omega \in \Omega' \cap B(\beta - \varepsilon)$, then there is a constant $K = K(\omega)$ such that (1.7) is satisfied with v/K in place of v and $\beta - \varepsilon$ in place of λ. Therefore, by Theorem 1.4 and Remark 6.1 we have

$$\overline{\lim_{h \downarrow 0}} \frac{w_{T-(t+h)} - w_{T-t}}{\sqrt{h}} \geq -\nu c(\beta - \varepsilon)$$

whenever $\omega \in \Omega' \cap B(\beta - \varepsilon)$ and $t \in [0,T)$. This proves assertion (i) of Theorem 1.5 with $c_0 = \nu c(\beta - \varepsilon)$.

To prove assertion (ii) notice that, due to the above, one can take

$$c_0 = c_1 := \inf\{\nu c(\beta - \varepsilon) : \sigma \in (0,1), r \geq 2, \beta > 0, \varepsilon \in (0, \beta)\}.$$

Since $c(\lambda)$ is a continuous decreasing function, we have

$$c_1 = \inf\{\nu c(\beta) : \sigma \in (0,1), r \geq 2, \beta > 0\}.$$

Next, if $0 < \sigma < 1$, then the largest value of β is

$$\beta_0 = 2\,\frac{(1-\sigma)^2}{2-\sigma^2},$$

which occurs for $r = (2-\sigma^2)(\sigma - \sigma^2)^{-1}$. Notice that in similar computations in Remark 3.7 of [8] there is a misprint in the expression of β_0, which contains $1 - \sigma^2$ in the numerator in [8] instead of $(1-\sigma)^2$. Thus,

$$c_1 = \inf_{\sigma \in (0,1)} c\left(2\,\frac{(1-\sigma)^2}{2-\sigma^2}\right)\sigma^{-1}\sqrt{2-\sigma^2}.$$

We give an estimate for this inf by using Lemma 4.1 and letting $\sigma \downarrow 0$. Then we see that (a.s.)

$$c_1 \leq \lim_{\sigma \downarrow 0} c\left(2\,\frac{(1-\sigma)^2}{2-\sigma^2}\right)\sigma^{-1}\sqrt{2-\sigma^2} = \sqrt{\pi}\lim_{\sigma \downarrow 0}\sigma^{-1}\left(1 - 2\,\frac{(1-\sigma)^2}{2-\sigma^2}\right) = 2\sqrt{\pi}.$$

This proves assertion (ii) and the theorem.

Remark 6.2. It turns out that taking $\sigma = 0.31$ yields a slightly better approximation of c_1, but still very far from the sharp value of c_0, which is 1.

Acknowledgments. The author is sincerely grateful to D. Aronson, N. Jain, W. Littman, E. Perkins, M. Safonov, H. Weinberger, and O. Zeitouni for fruitful discussions. This article took its almost final form after the author gave a talk at the conference on Stochastic Partial Differential Equations, Levico Terme (Trento), 2002, and the author uses this opportunity to thank the organizers of the conference for the invitation. Finally, he wishes to thank the referee for several very useful comments.

REFERENCES

[1] J.R. Cannon, *The One-Dimensional Heat Equation*, Encyclopedia Math. Appl. 23, Addison–Wesley, Reading, MA, 1984.

[2] B. Davis, *On Brownian slow points*, Z. Wahrsch. Verw. Gebiete, 64 (1983), pp. 359–367.

[3] E.B. Dynkin, *Markov Processes*, Vols. I, II, Grundlehren Math. Wiss. 121–122, Springer-Verlag, Berlin, 1965.

[4] L.C. Evans and R.F. Gariepy, *Wiener's criterion for the heat equation*, Arch. Ration. Mech. Anal., 78 (1982), pp. 293–314.

[5] K. Itô and H.P. McKean, *Diffusion Processes and Their Sample Paths*, Classics Math., Springer-Verlag, Berlin, 1974.

[6] N.V. Krylov, *On the regular boundary points for Markov processes*, Theory Probab. Appl., 11 (1966), pp. 609–614.

[7] N.V. Krylov, *Weighted Sobolev spaces and Laplace's equation and the heat equations in a half space*, Comm. Partial Differential Equations, 24 (1999), pp. 1611–1653.

[8] N.V. Krylov, *SPDEs in $L_q((0,\tau],L_p)$ spaces*, Electron. J. Probab., 5 (2000), paper 13; available from http://www.math.washington.edu/~ejpecp.

[9] N.V. Krylov, *Some properties of traces for stochastic and deterministic parabolic weighted Sobolev spaces*, J. Funct. Anal., 183 (2001), pp. 1–41.

[10] N.V. Krylov and S.V. Lototsky, *A Sobolev space theory of SPDEs with constant coefficients on a half line*, SIAM J. Math. Anal., 30 (1998), pp. 298–325.

[11] N.V. Krylov and S.V. Lototsky, *A Sobolev space theory of SPDEs with constant coefficients in a half space*, SIAM J. Math. Anal., 31 (1999), pp. 19–33.

[12] S.V. Lototsky, *Dirichlet problem for stochastic parabolic equations in smooth domains*, Stochastics Stochastics Rep., 68 (1999), pp. 145–175.

[13] E. Perkins, *On the Hausdorff dimension of the Brownian slow points*, Z. Wahrsch. Verw. Gebiete, 64 (1983), pp. 369–399.

[14] K. Uchiyama, *Brownian first exit from and sojourn over one-sided moving boundary and application*, Z. Wahrsch. Verw. Gebiete, 54 (1980), pp. 75–116.

SIAM J. MATH. ANAL.
Vol. 34, No. 5, pp. 1183–1206

CALIBRATION OF THE LOCAL VOLATILITY IN A GENERALIZED BLACK–SCHOLES MODEL USING TIKHONOV REGULARIZATION*

S. CRÉPEY†

Abstract. Following an approach introduced by Lagnado and Osher [*J. Comput. Finance*, 1 (1) (1997), pp. 13–25], we study Tikhonov regularization applied to an inverse problem important in mathematical finance, that of calibrating, in a generalized Black–Scholes model, a local volatility function from observed vanilla option prices.

We first establish $W_p^{1,2}$ estimates for the Black–Scholes and Dupire equations with measurable ingredients. Applying general results available in the theory of Tikhonov regularization for ill-posed nonlinear inverse problems, we then prove the stability of this approach, its convergence towards a minimum norm solution of the calibration problem (which we assume to exist), and discuss convergence rates issues.

Key words. options, calibration, ill-posed nonlinear inverse problem, Tikhonov regularization, parameter estimation, $W_p^{1,2}$ estimates

AMS subject classifications. 35K15, 35Q80, 35R05, 35R30

PII. S0036141001400202

1. Introduction.

A quantity of fundamental importance to the trading of options on a stock S is the stochastic component in the evolution of the stock price, the so-called *volatility*. Obtaining estimates for the volatility is a major challenge for market finance. Unlike historical estimates of the volatility, based upon observations of the time-series of the stock price, calibration estimates rely upon the anticipation of the trading agents reflected in the prices of the traded option products derived from S. We consider in this article Tikhonov regularization applied to a widely studied inverse problem in mathematical finance, that of calibrating a local volatility function from a given set of option prices in a generalized Black–Scholes model.

This calibration problem has received intensive study in the last ten years; see, for instance, [19, 17, 18, 35, 1, 11, 8, 34, 2, 29, 24, 7, 14, 15, 4] and references therein. Notable approaches include entropy regularization (Avellaneda et al. [2]) or parametrix expansion (Bouchouev and Isakov [8]). In this paper, we shall focus upon the Tikhonov regularization method, following an approach introduced by Lagnado and Osher [29]. Jackson, Süli, and Howison [24] devised an implementation of this method with splines. Bodurtha and Jermakyan used linearization [7]. However, while most previous approaches adopted a numerical and empirical point of view, our aim is to establish a rigorous theoretical ground for this inverse problem in a partial differential equation framework.

Work corresponding to a first stage of this research has been published in my Ph.D. thesis [14, Part IV] (in French), while a preliminary version of this article has been published as a CMAP Internal Research Report [15]. A further article addresses an implementation of the method in a trinomial tree (explicit finite differences) setting and reports numerical experiments illustrating the stability of the local volatility function thus calibrated [16].

*Received by the editors December 24, 2001; accepted for publication (in revised form) November 4, 2002; published electronically April 17, 2003.
http://www.siam.org/journals/sima/34-5/40020.html
†Artabel SA, 69 rue de Paris, 91400 Orsay, France (stephane.crepey@artabel.net).

2. Preliminaries. In this section, we will give an informal presentation of the calibration problem and of the Tikhonov regularization method, provide an overview of the paper, and define the main notation and general assumptions.

2.1. Generalized Black–Scholes model. In market finance, a European *call* (respectively, *put*) option with maturity date T and strike K, on an underlying asset S, denotes a right to buy (respectively, sell), at price K, a unit of S at time T. Let us then consider a theoretical financial market, with two traded assets: cash, with constant interest rate r, and a risky stock, with diffusion price process

$$dS_t = S_t(\rho(t,S_t)dt + \sigma(t,S_t)dW_t)\ , \quad t > t_0\ ; \quad S_{t_0} = S_0\ .$$

Here W means a standard Brownian motion. Moreover, the stock is assumed to yield a continuously compounded dividend at constant rate q. Suppose finally that the market is liquid, nonarbitrable, and perfect. These assumptions mean, respectively, that first, there are always buyers and sellers; second, there can be no opportunity that a riskless investment can earn more than the interest rate of the economy r; and third, there are no restrictions of any kind on the sales and no transaction costs. Under these assumptions the market is complete. This means that any option can be duplicated by a portfolio of cash and stock. Moreover, a European call/put on S has a theoretical fair price within the model, which we will denote by $\Pi^{+/-}_{T,K}(t_0,S_0;r,q,a)$, where $a \equiv \sigma^2/2$, and

$$\Pi^{+/-}_{T,K}(t_0,S_0;r,q,a) = e^{-r(T-t_0)}\mathrm{E}_P^{t_0,S_0}(S_T-K)^{+/-}\ . \tag{2.1}$$

Here P denotes the so-called *risk-neutral* probability, under which

$$dS_t = S_t((r-q)dt + \sigma(t,S_t)dW_t)\ , \quad t > t_0\ ; \quad S_{t_0} = S_0\ . \tag{2.2}$$

Alternatively to the probabilistic representation (2.1), the prices $\Pi^{+/-}$ can be given as the solution to a differential equation. One can use either the *Black–Scholes* backward parabolic equation in the variables (t_0,S_0), which is

$$\left\{\begin{array}{l} -\partial_t\Pi - (r-q)S\partial_S\Pi - a(t,S)S^2\partial^2_{S^2}\Pi + r\Pi \ = \ 0, \quad t<T, \\ \Pi|_T \ \equiv \ (S-K)^{+/-}\ , \end{array}\right. \tag{2.3}$$

or the *Dupire* forward parabolic equation, in the variables (T,K), given by

$$\left\{\begin{array}{l} \partial_T\Pi - (q-r)K\partial_K\Pi - a(T,K)K^2\partial^2_{K^2}\Pi + q\Pi \ = \ 0, \quad T>t_0, \\ \Pi|_{t_0} \ \equiv \ (S_0-K)^{+/-}\ . \end{array}\right. \tag{2.4}$$

We will show in Lemma 4.1 and Theorem 4.3 that (2.1) or (2.3)–(2.4) hold for an arbitrary measurable, positively bounded local volatility function a. However, let us give a less formal insight by recalling the Black–Scholes seminal analysis [6], valid in the special case where the volatility depends on time alone. We consider a self-financing portfolio, short one option and long $\partial_S\Pi$ shares of the underlying stock. The value V of the risky component of the portfolio then evolves as

$$\begin{aligned} dV_t &= -d\Pi(t,S_t) + \partial_S\Pi(dS_t + qS_tdt) \\ &= -(\partial_t\Pi - qS\partial_S\Pi + aS^2\partial^2_{S^2}\Pi)dt \end{aligned}$$

from Itô's lemma. Since V has a deterministic rate of return, absence of opportunity of arbitrage implies that this rate equals the riskless interest rate r. Otherwise said,

$$-\partial_t\Pi + qS\partial_S\Pi - aS^2\partial^2_{S^2}\Pi = r(-\Pi + S\partial_S\Pi)\ ,$$

whence (2.3). As for (2.1), it can be viewed as the Feynman–Kac representation for the solution of (2.3). Notice that this analysis does not rely on the specific character of the payoff of the call or put option. However, the opposite is true for (2.4). It is indeed, as noticed by Dupire [19], a Fokker–Planck equation integrated twice with respect to the space variable K, using moreover the formal identity

$$\partial^2_{K^2}(S_0 - K)^{+/-} \equiv \delta_{S_0}(K) \ ,$$

where δ_{S_0} denotes the Dirac mass at S_0.

2.2. Direct and inverse problems. In the special case where the volatility, $a \equiv \sigma^2/2$, is a constant, or a function of time alone, explicit formulas for the prices $\Pi^{+/-}$ are known (see Black and Scholes [6] or Merton [31]). But in the case of a general local volatility function $a(t, S)$, one must turn to finite differences or a Monte Carlo procedure based upon (2.3)–(2.4) or (2.1). Moreover, observation teaches that no constant or merely time-dependent local volatility function is consistent with most sets of market quotes. This phenomenon is known by market practitioners as the *smile of implied volatility*.

However, in practice it is not the local volatility that is known but the prices themselves. In fact the local volatility is the only quantity in (2.1) or (2.3)–(2.4) which cannot be obtained from the market. Indeed r and q, as well as, to some extent, Π, can all be retrieved from market-quoted quantities. Consequently, one usually wishes to solve the inverse problem: finding $a(t, S)$ such that the theoretical prices given by (2.1) or (2.3)–(2.4) match the observed option prices. We thus use liquid quotations of actively traded options, which are usually referred to as *vanilla* options, as a way to extract information about the future behavior of the underlying asset. The calibrated local volatility function is then used by risk managers or traders to value risk exposure, or price *exotic* (nonvanilla) options and calculate hedge ratios consistently with the market.

This is the problem we will be concerned with here. In particular, there are two cases which are commonly considered in the literature, and we will treat both in parallel. In the first one, this matching is required to occur on the actual, hence finite, set of pairs (T, K) with observed prices. In the second case, the matching is required to occur over all (T, K) such that $T \geq t_0, \ K > 0$. This makes sense, for example, if the actual set of observed prices has been interpolated. To distinguish between these two cases, we will refer to the first as the *discrete* calibration problem and the second as the *continuous* calibration problem.

2.3. The Tikhonov regularization method. Both the discrete and continuous calibration problems are ill-posed. This is the case in the continuous calibration problem because the solution depends upon the data in an unstable manner, and in the discrete calibration problem because the full surface $a(t, S)$ is simply underdetermined by the discrete data. It is then necessary to introduce stabilizing procedures in the reconstruction method for the local volatility function. One of these is known as the Tikhonov regularization method [38, 21]. The idea is to tackle the calibration problem as a minimization problem, where the cost criterion to be minimized is

$$J_\alpha(a) \equiv d(\Pi(a), \pi)^2 + \alpha\rho(a, a_0)^2 \ .$$

Here $d(\Pi(a), \pi)$ denotes a distance between the model prices $\Pi(a)$ and the observed prices π, α is the regularization parameter, and ρ is a penalty designed to keep a

close to the so-called *prior* a_0, which reflects a priori information about a. Following Lagnado and Osher [29], we shall choose $\rho(a, a_0)^2 \equiv \|a - a_0\|^2_{H^1}$, where

$$\|u\|^2_{H^1} \equiv \int\int u^2 + \|\nabla u\|^2,$$

which is the H^1-(squared) norm of u with logarithmic variables $t, y = \ln(S)$.

2.4. Overview. We first study, in an appropriate functional analysis setting, Black–Scholes and Dupire linear parabolic equations with measurable ingredients (sections 3 and 4). These are linear one-dimensional equations in nondivergence form, with positively bounded dominant coefficients. We thus extend well-known results when the dominant coefficient a is a regular function. Mixing the probabilistic pointwise and L_p estimates of Krylov [26] with the analytic $W^{1,2}_p$ estimates of Fabes [23] and Stroock and Varadhan [37], we obtain $W^{1,2}_p$ estimates for the equations with source terms. Using the theory of L_p-viscosity solutions [10, 13], we then show that our equations admit unique solutions, for which we provide a probabilistic representation (Theorems 4.2 and 4.3).

Proposition 5.1 sums up the main properties of the pricing functional Π useful for the study of the calibration problems, namely, compactness, twice Gâteaux differentiability and stability with respect to perturbations of parameters. We can then apply the general theory of Tikhonov regularization for ill-posed nonlinear inverse problems [21, 22, 27, 32, 33] to both the continuous and discrete calibration problems. We thus prove the stability of the method for arbitrary values of the regularization parameter (section 5). Assuming the existence of a solution of the calibration problem, we prove the convergence of the method towards an a_0-minimum norm solution when the regularization parameter tends to 0, and we exhibit conditions sufficient to ensure convergence rates in $\mathrm{O}(\sqrt{\delta})$, where δ is the data noise (section 6).

2.5. Main notation and general assumptions. To avoid much repetition, we define now a set of notation and related general assumptions that will be assumed to hold throughout the paper. When stronger assumptions are required, they will be stated explicitly in the body of the paper.

General notation.

- $x \wedge y$, $x \vee y$: $\min(x, y)$, $\max(x, y)$.
- x^+, x^-: $\max(x, 0)$, $\max(-x, 0)$.
- $C, C', \ldots, C \equiv C_\rho(\rho_1, \ldots, \rho_n)$: Constants $C, C', \ldots$ depending upon nothing but the parameters $\rho, \rho_1, \ldots, \rho_n$.

One should be aware that these constants may vary with the context. We will also use the notation "$\equiv$" for "denotes" or "equals identically" (that is, equality between functions), according to the context.

Mathematical finance.

- S, $y = \ln(S)$: Lognormal underlying diffusion in financial and logarithmic variables.
- $q, r \in [0, R]$: Dividend yield attached to S, short rate of the economy.
- $a \equiv \sigma^2/2$, a_0: Local volatility function, prior a_0 on a.
- $\underline{a}$, $\overline{a}$, $\hat{a}$: Bounds on a_0 and a such that $0 < \underline{a} < \overline{a}$, $\hat{a} \equiv (\underline{a} + \overline{a})/2$.
- $\overline{p} \equiv \overline{p}(\underline{a}, \overline{a})$: A real in $]2, 3[$ depending upon $\underline{a}$ and $\overline{a}$; see Theorem 4.2.
- W: Standard Brownian motion.
- $Q =]\underline{t}, \overline{T}[\times \mathbb{R}$: A plane strip on which a is defined, in logarithmic variables.

- $(t_0, y_0), (T, k)$: Points in $\overline{Q}$, with $t_0 \leq T$.
- $\overline{y}_0, \overline{k}$: Bounds on $|y_0|$, $|k|$.
- Q_{t_0}, Q^T: $Q \cap \{t > t_0\}$, $Q \cap \{t < T\}$.
- $\overline{Q}_{t_0}$, $\overline{Q}^T$: Closures of Q_{t_0}, Q^T.
- $\Pi^{+/-}_{T,K}(t_0, S_0; r, q, a)$, $\Pi^{+/-}_{T,k}(t_0, y_0; r, q, a)$: The price, in a generalized Black–Scholes model, for a European call/put option with maturity T and strike $K = e^k$, at the current phase $t_0, S_0 = e^{y_0}$, in financial and logarithmic variables.
- $\gamma_{t_0,y_0}(t, y; r, q, a)$: Transition probability density discounted at rate r (that is, $e^{-r(t-t_0)}\times$ the density), for the underlying diffusion in logarithmic variable y.
- $BS^{+/-}_{Q^T}(k; r, q, a)$, $BS'_{Q^T}(r, q, a; \Gamma)$, $DUP^{+/-}_{Q_{t_0}}(y_0; r, q, a)$: Black–Scholes call/put equation on Q^T, Black–Scholes derived equation with source term Γ, Dupire call/put equation on Q_{t_0}; see section 3.2.

To alleviate notation, r, q, a will sometimes be abbreviated to a; $\Pi^{+/-}_{T,K}(t_0, S_0; a)$ or $\Pi^{+/-}_{T,k}(t_0, y_0; a)$ to $\Pi^{+/-}$; $BS^{+/-}_{Q^T}(k; a)$, $BS'_{Q^T}(a; \Gamma)$, $DUP^{+/-}_{Q_{t_0}}(y_0; a)$, and $\gamma_{t_0,y_0}(t, y; a)$ to $BS^{+/-}$, BS', $DUP^{+/-}$, and γ, respectively.

In the case of the call option, we will sometimes drop the $^+$ superscript. For instance, by default, Π will refer to Π^+.

Functional analysis.

- Ω: Regular by parts, open plane area.
- p, θ: Real $p \in]2, +\infty[$, $\theta \equiv 1 - 2/p > 0$.
- $L_p(\Omega)$, $L_{p,loc}(\Omega)$, $H^1(\Omega)$, $H^2(\Omega)$, $W^1_p(\Omega)$, $W^{1,2}_p(\Omega)$, $W^{1,2}_{p,loc}(\Omega)$, $\mathcal{C}^0_\theta(\overline{\Omega})$, $\mathcal{D}(\overline{\Omega})$: Sobolev spaces on Ω; see section 3.1.
- Γ: Element of $L_p(Q)$.
- $\mathcal{M}_Q(\underline{a}, \overline{a})$: Set of real measurable functions on Q with bounds $\underline{a}$ and $\overline{a}$.
- $a_0 + H^1_Q(\underline{a}, \overline{a})$: Set of functions in $a_0 + H^1(Q)$ with bounds $\underline{a}$ and $\overline{a}$.
- h, h': Elements of $H^1(Q)$.
- $\mathcal{E} \rightarrow$: Convergence in the topology of the space $\mathcal{E}$.
- $\|X\|$, $\|X\|_{\mathcal{E}}$: Euclidean norm of X, norm of X in the surrounding normed space $\mathcal{E}$.
- $\langle X, Y\rangle$, $\langle X, Y\rangle_{\mathcal{E}}$: Inner product of X and Y in the surrounding Euclidean space, Hilbert space $\mathcal{E}$.
- $d\Pi(a).h$: Derivative in the direction h of the functional Π at the local volatility function a.
- $d\Pi(a)^\star$: Adjoint operator of $h \mapsto d\Pi(a).h$; see section 6.2.
- $\nabla J(a)$: Gâteaux derivative of the cost criterion J at the local volatility function a.

For instance, if J denotes a cost criterion on a Hilbert space $\mathcal{E}$, then in our notation

$$\langle \nabla J(a)\ , h\rangle_{\mathcal{E}} \ = \ dJ(a).h\ , \quad h \in \mathcal{E}\ .$$

In the same way, the general assumptions we have made above on a and a_0 can be stated as

$$a_0, a \in \mathcal{M}_Q(\underline{a}, \overline{a})\ .$$

Finally, we will refer to the statements in Remark 3.5 and Lemma 4.1(3) as *symmetry* and *parity*, respectively.

3. Strong solutions of parabolic problems.

3.1. Functional spaces and Sobolev embeddings. Let us first introduce some Hilbert and Banach spaces, which we will use as spaces of local volatility functions and solutions of Black–Scholes and Dupire equations.

Given the open plane area Ω, we will denote by $\mathcal{D}(\overline{\Omega})$ the space of traces on Ω of regular functions with compact support in the plane. We will use the usual Hilbert spaces $H^2(\Omega) \subset H^1(\Omega) \subset L_2(\Omega)$ and the Banach spaces $\mathcal{C}^0_\theta(\overline{\Omega})$, $L_p(\Omega)$, $W^1_p(\Omega)$, $W^{1,2}_p(\Omega)$, where

$$\begin{aligned}
\|u\|_{\mathcal{C}^0_\theta(\overline{\Omega})} &= \sup_{(t,y)\in\overline{\Omega}} |u| \;+\; \sup_{(t,y)\neq(t',y')\in\overline{\Omega}} \frac{|u(t,y)-u(t',y')|}{|t-t'|^\theta + |y-y'|^\theta};\\
\|u\|_{W^1_p(\Omega)} &= \|u\|_{L_p(\Omega)} \;+\; \|\partial_t u\|_{L_p(\Omega)} \;+\; \|\partial_y u\|_{L_p(\Omega)};\\
\|u\|_{W^{1,2}_p(\Omega)} &= \|u\|_{L_p(\Omega)} \;+\; \|\partial_t u\|_{L_p(\Omega)} \;+\; \|\partial_y u\|_{L_p(\Omega)} \;+\; \|\partial^2_{y^2} u\|_{L_p(\Omega)}\,.
\end{aligned}$$

Finally, we will denote by $W^{1,2}_{p,loc}(\Omega)$ the localized Fréchet space of functions which belong to $W^{1,2}_p(\Omega')$ for every regular open bounded subset Ω' with $\overline{\Omega}' \subset \Omega$.

Now we have the following Sobolev embeddings, for which the reader is referred, for instance, to Larrouturou and Lions [30]:

1. For Ω bounded or half-plane,

$$W^1_p(\Omega) \hookrightarrow \mathcal{C}^0_\theta(\overline{\Omega})\,. \tag{3.1}$$

This embedding notably implies the existence of a unique continuous extension up to the boundary for the strong solutions introduced by item 1 of Definition 3.1 below.

2. For Ω bounded,

$$H^1(\Omega) \hookrightarrow L_p(\Omega)\,. \tag{3.2}$$

This embedding, called the Rellich–Kondrakov embedding, is *compact*, which means that it maps weakly convergent sequences into strongly convergent ones.

Let us now present the definitions of a solution of a partial differential equation that we will need. For more about these definitions, the reader is referred to Ladyzhenskaya, Solonnikov, and Ural'tseva [28], Crandall, Kocan, and Swiech [13], Wang [39], Caffarelli et al. [10], and Crandall, Ishii, and Lions [12].

DEFINITION 3.1. *Let there be a linear parabolic equation on $\overline{\Omega}$, with measurable ingredients and a continuous boundary condition on $\partial_p\Omega$, the parabolic boundary of Ω.*

1. *We call a function a* strong solution in $L_p(\Omega)$*, or an* $L_p(\Omega)$-solution*, if it is a function in $W^{1,2}_p(\Omega)$, which satisfies the boundary condition, and solves the equation almost everywhere. We also use this definition with $W^{1,2}_{p,loc}(\Omega)$ to define a* strong solution in $L_{p,loc}(\Omega)$*, or an* $L_{p,loc}(\Omega)$-solution.

2. *We call a function an* $L_{p,loc}(\Omega)$-viscosity solution *if it is a continuous function on $\overline{\Omega}$, which satisfies the boundary condition, and solves the equation in the viscosity meaning for test functions in $W^{1,2}_{p,loc}(\Omega)$.*

The relations between these definitions of a solution are as follows (see Crandall, Kocan, and Swiech [13]):

1. An $L_{p,loc}(\Omega)$-solution is an $L_{p,loc}(\Omega)$-viscosity solution.

2. Conversely, an $L_{p,loc}(\Omega)$-viscosity solution that belongs to $W^{1,2}_{p,loc}(\Omega)$ is an $L_{p,loc}(\Omega)$-solution.

The following theorem gathers the main properties of the Sobolev spaces on plane strips that we will need.

THEOREM 3.2.

1. *$H^1(Q)$ is continuously embedded in $L_p(Q)$.*
2. *$\mathcal{D}(\overline{Q})$ is dense in $L_p(Q)$, $H^1(Q)$, $H^2(Q)$.*
3. *The application*

$$\mathcal{D}(\overline{Q}) \times \mathcal{D}(\overline{Q}) \ni (u, v) \mapsto (u|_{\partial Q}, \partial_n v) \in L_2(\partial Q)^2 \ ,$$

where $\partial_n v$ denotes the normal derivative, admits a unique linear continuous extension, called trace*, from $H^1(Q) \times H^2(Q)$ to $L_2(\partial Q)^2$.*

4. *The set of traces on ∂Q of functions of $H^1(Q) \times H^2(Q)$ forms a dense subset of $L^2(\partial Q)^2$, and we have the so-called* generalized Green formula *for every $(u, v) \in H^1(Q) \times H^2(Q)$:*

$$-\int\int_Q u\ (\Delta v) \ = \ \int\int_Q \langle \nabla u, \nabla v \rangle \ - \ \int_{\partial Q} u\ \partial_n v \ .$$

Proof. These properties result from the analogous properties well known on open half-planes (see, for instance, Larrouturou and P. L. Lions [30], Bensoussan and J.-L. Lions [3]). For details, the reader is referred to Crépey [14, Theorem F.1] and the proof given therein. □

In the upcoming proofs, we will often be able to proceed by density thanks to the following lemma.

LEMMA 3.3. *There exist Lipschitzian functions $a_n \in \mathcal{M}_Q(\underline{a}, \overline{a})$ ($n \in \mathbb{N}^\star$) such that a_n converges to a in $L_{p,loc}(Q)$ when $n \to +\infty$.*

Proof. This follows from standard mollification with compact support, applied to a extended by zero outside Q (see, for instance, Brézis [9]). □

3.2. Black–Scholes, Dupire, and derived equations. Let us now introduce the main equations in this work.

DEFINITION 3.4.

1. *We define the Black–Scholes call/put equation, $BS^{+/-}_{Q^T}(k; r, q, a)$, with backward logarithmic variables $(t, y) \in \overline{Q}^T$, parameterized by (T, k), as*

$$\left\{ \begin{array}{l} -\partial_t \Pi - (r - q - a(t,y))\, \partial_y \Pi - a(t,y) \partial^2_{y^2} \Pi + r\Pi \ = \ 0 \quad \text{on } Q_T \ , \\ \Pi|_T = (e^y - e^k)^{+/-} \ . \end{array} \right.$$

We also define the Black–Scholes derived equation with source term Γ, $BS'_{Q^T}(r, q, a; \Gamma)$, as

$$\left\{ \begin{array}{l} -\partial_t (\delta\Pi) - (r - q - a(t,y))\, \partial_y (\delta\Pi) - a(t,y) \partial^2_{y^2} (\delta\Pi) + r(\delta\Pi) \ = \ \Gamma \quad \text{on } Q_T \ , \\ \delta\Pi|_T \ \equiv \ 0 \ . \end{array} \right.$$

2. *We define the Dupire call/put equation, $DUP^{+/-}_{Q_{t_0}}(y_0; r, q, a)$, with forward logarithmic variables (T, k), at the current phase (t_0, y_0), as*

$$\left\{ \begin{array}{l} \partial_T \Pi_{T,k} - (q - r - a(T,k)) \partial_k \Pi_{T,k} - a(T,k) \partial^2_{k^2} \Pi_{T,k} + q \Pi_{T,k} \ = \ 0 \quad \text{on } Q_{t_0} \ , \\ \Pi|_{t_0} \ \equiv \ (e^{y_0} - e^k)^{+/-} \ . \end{array} \right.$$

3. *Finally, we define the diffusion underlying the previous problems, with logarithmic variables, as*

$$dy_t \;=\; \left(r-q-\frac{\sigma(t,y_t)^2}{2}\right)\,dt+\sigma(t,y_t)\;dW_t\;,\qquad y_{t_0}\;=\;y_0\;. \tag{3.3}$$

Remark 3.5 (symmetry). Changing, moreover, the direction of time T, via $\tau\equiv\overline{T}+t_0-T$, $\check{\phi}(\tau,k)\equiv\phi(T,k)$ for any function ϕ, then $DUP^{+/-}_{Q_{t_0}}(y_0;r,q,a)$ becomes $BS^{-/+}_{Q_{t_0}}(y_0;q,r,\check{a})$.

LEMMA 3.6.

1. *(Black–Scholes and Dupire equations.) Equations $BS^{+/-}$ have at most one $L_{p,loc}(Q^T)$-solution Π such that $|\Pi|\le K\vee S$.*

2. *(Derived equations.) For any $L_p(Q^T)$-solution $\delta\Pi$ of BS', we have*

$$\|\delta\Pi\|_{\mathcal{C}^0_\theta(\overline{Q}^T)}\;\le\;C'\;\|\delta\Pi\|_{W^{1,2}_p(Q^T)}\;, \tag{3.4}$$

where $C'\equiv C'_p$. Moreover, $\delta\Pi$ is also the unique $L_{p,loc}(Q^T)$-solution of BS' which converges to 0 when $|y|\to+\infty$, uniformly with t.

Proof. 1. Given two such solutions Π and Π', let us define $\delta\Pi\equiv e^{-2y+\rho t}(\Pi-\Pi')$, where $\rho=r+2\overline{a}$. By linearity, $\delta\Pi$ is an $L_{p,loc}(Q^T)$-solution of

$$\left\{\begin{array}{l}-\partial_t\delta\Pi-(r-q+3a)\,\partial_y\delta\Pi-a\partial^2_{y^2}\delta\Pi+(2q+2\overline{a}-2a)\delta\Pi=0,\\ \delta\Pi|_T\;\equiv\;0\;.\end{array}\right. \tag{3.5}$$

Moreover, let us fix $\varepsilon>0$. One can choose $Y_\varepsilon\ge1/\varepsilon$ such that for $|y|\ge Y_\varepsilon$, we have $|\delta\Pi(t,y)|\le 2e^{-2y+\rho t}(K\vee e^y)\le\varepsilon$, uniformly with $t\in[\underline{t},T]$. Then $|\delta\Pi|\le\varepsilon$ on $Q^T\cap\{|y|\le Y_\varepsilon\}$, by the maximum principle in Crandall, Kocan, and Swiech [13, Proposition 2.6]. So $\delta\Pi\equiv0$ on Q^T by passage to the limit when $\varepsilon\to0$.

2. By the same maximum principle as above, we have uniqueness in the class of $L_{p,loc}(Q^T)$-solutions of BS' which converge to 0 when $|y|\to+\infty$, uniformly with t. Now, let us be given an $L_p(Q^T)$-solution $\delta\Pi$ of BS'. Since the solution $\delta\Pi$ is continuous on $\overline{Q}^T$ and vanishes at T, it may be identified with an element of $W^1_p(\Omega)$, where $\Omega\equiv\,]\underline{t},+\infty[\times\mathbb{R}$, by extension with 0 on the right of T. Estimate (3.4) then follows from the Sobolev embedding (3.1) on the half-plane Ω. Finally, $\delta\Pi\in\mathcal{C}^0_\theta(\overline{Q}^T)\cap L_p(Q^T)$ converges to 0 when $|y|\to+\infty$, uniformly with t. ∎

4. Existence, uniqueness, and probabilistic representation of solutions.

4.1. Diffusion. The following lemma links the price of a European call/put with the discounted expectation of the corresponding payoff in a generalized Black–Scholes model.

LEMMA 4.1.

1. *The diffusion equation (2.2) has a unique weak solution on $]t_0,\overline{T}[$:*

$$S_t\;=\;S_0\;e^{(r-q)(t-t_0)}\;\exp\left(\int_{t_0}^t\sigma(s,S_s)dW_s-\frac{1}{2}\int_{t_0}^t\sigma^2(s,S_s)ds\right)\;,\quad t\in\,]t_0,\overline{T}[,$$

where the last exponential is a martingale, under the risk-neutral probability P. In particular,

$$E_P^{t_0,S_0}S_t=S_0e^{(r-q)(t-t_0)}\;,\quad t\in\,]t_0,\overline{T}[\;. \tag{4.1}$$

2. *The price* $\Pi^{+/-}$ *equals the payoff expectation of the call/put at* T*, discounted at rate* r*:*

$$\Pi^{+/-} = e^{-r(T-t_0)} E_P^{t_0,S_0} \left(S_T - K\right)^{+/-}$$

under the risk-neutral probability P*. In particular,* $0 \le \Pi \le S_0$*.*

3. *Denoting* $\Pi^+ - \Pi^-$ *by* $\delta\Pi$*, we have*

$$\delta\Pi \equiv S_0 e^{-q(T-t_0)} - K e^{-r(T-t_0)} \ .$$

This relation, known as call/put parity*, notably implies that* $\partial^2_{S^2}\delta\Pi$*,* $\partial^2_{K^2}\delta\Pi$*,* $(\partial^2_{y^2} - \partial_y)\delta\Pi$*, and* $(\partial^2_{k^2} - \partial_k)\delta\Pi$ *all vanish identically.*

Proof. 1. For the proof, see, for instance, Stroock and Varadhan [37, Exercise 7.3.3] and Karatzas and Shreve [25, Problem 5.6.15 and Corollary 3.5.13].

2. and 3. The expression for $\Pi^{+/-}$ then follows from Karatzas and Shreve [25, section 5.8.A]. Using this expression, the bounds on Π and the call/put parity proceed from (4.1). ∎

4.2. Derived hedge equations with source terms. The following theorem and the estimate (4.3) therein are the cornerstones of this article. The difficulty comes from the lack of regularity of the local volatility function a, which is merely required to be measurable and positively bounded. But this turns out to be sufficient in the present one-dimensional linear framework. Recall that Γ denotes an element of $L_p(Q)$.

THEOREM 4.2. *There exists* $\overline{p} \equiv \overline{p}(\underline{a}, \overline{a}) \in \,]2,3[$ *such that if* $p \in \,]2, \overline{p}[$*, then, when* (t,y) *varies within* $\overline{Q}^T$*,*

$$\delta\Pi(t,y) \ = \ E_P^{t,y} \int_{s=t}^{T} e^{-r(s-t)} \Gamma(s, y_s)\ ds \tag{4.2}$$

is the only $L_p(Q^T)$*-solution, or* $L_{p,loc}(Q^T)$*-solution converging to* 0 *when* $|y| \to +\infty$*, uniformly with* t*, of* $BS'_{Q^T}(a;\Gamma)$*.*

Moreover,

$$\|\delta\Pi\|_{\mathcal{C}^0_\theta(\overline{Q}^T)} \ \le \ C' \ \|\delta\Pi\|_{W_p^{1,2}(Q^T)} \ \le \ C'C \ \|\Gamma\|_{L_p(Q^T)} \ , \tag{4.3}$$

where $C' \equiv C'_p$ *is as in* (3.4)*, and* $C \equiv C_p(\underline{t}, \overline{T}; R, \underline{a}, \overline{a})$*.*

Proof. For the moment, $p \in \,]2, +\infty[$. We first show that for $\varphi \in W_p^{1,2}(Q^T)$,

$$\|\varphi\|_{W_p^{0,1}(Q^T)} \ \le \ C_p \ \|\varphi\|^{1/2}_{W_p^{0,2}(Q^T)} \ \|\varphi\|^{1/2}_{L_p(Q^T)} \ . \tag{4.4}$$

Inequality (4.4) can be more readily seen on the following equivalent norms:

$$\|\varphi\|^p_{\widetilde{W}_p^{0,j}(Q^T)} \ \equiv \ \sum_{k \le j} \|\partial^k_{y^k}\varphi\|^p_{L_p(Q^T)} \ , \quad 0 \le j \le 2 \ .$$

Indeed, by integration over time of a classic Sobolev inequality (see, for instance, Bensoussan and J.-L. Lions [3, Chapter 2, equation (5.8)]):

$$\|\varphi\|^p_{\widetilde{W}_p^{0,1}(Q^T)} = \int_{t=\underline{t}}^{T} \|\varphi(t,\cdot)\|^p_{\widetilde{W}_p^1(\mathbb{R})} \ dt$$

$$\begin{aligned}
&\le C_p^p \int_{t=\underline{t}}^T \|\varphi(t,\cdot)\|^{p/2}_{\widetilde{W}^2_p(\mathbb{R})} \, \|\varphi(t,\cdot)\|^{p/2}_{L_p(\mathbb{R})} \, dt \\
&\le C_p^p \left(\int_{t=\underline{t}}^T \|\varphi(t,\cdot)\|^{p}_{\widetilde{W}^2_p(\mathbb{R})} \, dt \right)^{1/2} \left(\int_{t=\underline{t}}^T \|\varphi(t,\cdot)\|^{p}_{L_p(\mathbb{R})} \, dt \right)^{1/2} \\
&= C_p^p \, \|\varphi\|^{p/2}_{\widetilde{W}^{0,2}_p(Q^T)} \, \|\varphi\|^{p/2}_{L_p(Q^T)}
\end{aligned}$$

by the Cauchy–Schwarz inequality. This shows (4.4), which in turn implies

$$\|\varphi\|_{W^{0,1}_p(Q^T)} \;\le\; rC_p \|\varphi\|_{W^{0,2}_p(Q^T)} + C_p(r) \|\varphi\|_{L_p(Q^T)} \tag{4.5}$$

for every fixed $r > 0$, provided $C_p(r) \le C_p/4r$.

On the other hand, since (3.3) admits a unique weak solution (see item 1 of Lemma 4.1), then from Krylov [26, Theorem 2.4.5.a (proof) and Theorem 2.4.1]

$$E_P^{t,y} \int_t^T e^{-r(s-t)} |\Gamma(s, y_s)| \, ds \le C \, \|\Gamma\|_{L_p(Q^T)} \, , \tag{4.6}$$

where $C \equiv C_p(\underline{t}, T, R, \underline{a}, \overline{a})$.

We now assume that φ is an $L_p(Q^T)$-solution of $BS'_{Q^T}(a; \Gamma)$. For $\varepsilon > 0$, let τ_ε denote the exit time of $Q^T \cap \{|y| \le 1/\varepsilon\}$ for the y-process (3.3). It can be shown that (4.6) implies the following probabilistic representation:

$$E_P^{t,y} e^{-r(\tau_\varepsilon - t)} \varphi(\tau_\varepsilon, y_{\tau_\varepsilon}) - \varphi(t,y) \;=\; - E_P^{t,y} \int_{s=t}^{\tau_\varepsilon} e^{-r(s-t)} \Gamma(s, y_s) \, ds \, . \tag{4.7}$$

This has been shown by Bensoussan and J.-L. Lions [3, Chapter 2, section 8.3] in a variational context. We do not reproduce the proof here, though it proceeds in a similar fashion, using regularization and Itô's classic formula.

When $\varepsilon \to 0$, τ_ε almost surely converges to T. Moreover, φ is bounded and continuous. Estimate (4.6) then implies, through dominated convergence on the left- and right-hand sides of (4.7),

$$\varphi(t,y) = E_P^{t,y} \int_{s=t}^T e^{-r(s-t)} \Gamma(s, y_s) \, ds. \tag{4.8}$$

Then, from Krylov [26, Theorem 2.4.5.a],

$$\|\varphi\|_{L_p(Q^T)} \;\le\; C \, \|\Gamma\|_{L_p(Q^T)} \, , \tag{4.9}$$

where $C \equiv C_p(\underline{t}, T, R, \underline{a}, \overline{a})$. The probabilistic representation (4.8), for any a priori $L_p(Q^T)$-solution φ of $BS'_{Q^T}(a; \Gamma)$, also shows the consistency of such a priori solutions across various values of $p > 2$.

Moreover, by linearity, such an a priori solution φ is the $L_p(Q^T)$-solution of the equation $-\partial_t \varphi - \hat{a} \partial^2_{y^2} \varphi \;=\; \hat{\Gamma}$, where

$$\hat{\Gamma} \;=\; \Gamma - r\varphi + (r - q - a(t,y)) \partial_y \varphi + (a - \hat{a}) \partial^2_{y^2} \varphi \, ,$$

with homogeneous terminal condition. Therefore, following Stroock and Varadhan [37, Exercise 7.3.3, p. 211], we have the following estimate:

$$\begin{aligned}
&\|\partial^2_{y^2} \varphi\|_{L_p(Q^T)} \;\le\; C_p(\hat{a}) \\
&\times \left(\|\Gamma\|_{L_p(Q^T)} + R \|\varphi\|_{L_p(Q^T)} + (R + \overline{a}) \|\partial_y \varphi\|_{L_p(Q^T)} + \frac{1}{2} (\overline{a} - \underline{a}) \|\partial^2_{y^2} \varphi\|_{L_p(Q^T)} \right),
\end{aligned} \tag{4.10}$$

where $C_p(\hat{a})$ is a log-convex, hence continuous, function of $1/p$, also defined at $p = 2$, such that

$$C_{p=2}(\hat{a}) = \frac{1}{\hat{a}} < \frac{2}{\underline{a}}.$$

Therefore one can choose $\overline{p} \equiv \overline{p}(\underline{a}, \overline{a}) \in]2, 3[$ such that $C_p(\hat{a}) \leq \frac{2}{\underline{a}}$ if $p \in]2, \overline{p}[$. Estimate (4.3), at least with T instead of $\overline{T}$ in C, then results from (4.10), (4.5), (4.9), and (3.4). We will refer to the estimate (4.3) with T instead of $\overline{T}$ in C as *the temporary version* of estimate (4.3).

We now show the existence of an $L_p(Q^T)$-solution φ of $BS'_{Q^T}(a; \Gamma)$ in the special case where $\Gamma \in \mathcal{D}(\overline{Q}^T)$ by density using Lemma 3.3: Define $p' \equiv (2 + p)/2$. Following Fabes [23], $BS'_{Q^T}(a_n; \Gamma)$ admits an $L_p(Q^T) \cap L_{p'}(Q^T)$-solution φ_n. By the temporary version of estimate (4.3) and by successive extractions, one can find a subsequence $\varphi_{n'}$ that converges to a limit φ, weakly in $W_p^{1,2}(Q^T)$ or $W_{p'}^{1,2}(Q^T)$ and locally uniformly on $\overline{Q}^T$. By $W_p^{1,2}(Q^T)$-weak passage to the limit, φ inherits the temporary version of estimate (4.3). Then φ is an $L_p(Q^T)$-solution of $BS'_{Q^T}(a; \Gamma)$ by Lemma A.1. The general case where $\Gamma \in L_p(Q^T)$ follows straightaway by density using item 2 of Theorem 3.2.

Let us now consider the $L_p(Q)$-solution $\tilde{\varphi}$ of $BS'_Q(a; \widetilde{\Gamma})$, where $\widetilde{\Gamma} \equiv \Gamma/0$ on the left/right of T. By linearity and uniqueness of solutions of BS', $\tilde{\varphi}$ vanishes on Q_T, and $\tilde{\varphi}$ is equal to φ on Q^T. Therefore, the estimate (4.3) for φ on Q^T results from the temporary version of estimate (4.3) for $\tilde{\varphi}$ on Q. □

4.3. Homogeneous valuation equations. The following theorem is formally well known. When the local volatility function a is Hölderian (with logarithmic variables), it has indeed been justified by many authors. For instance, Dupire [19] and Bouchouev and Isakov [8] used partial differential equation arguments involving fundamental solutions. Alternatively, El Karoui [20] and Crépey [14, section 4.1, Part IV] used probabilistic arguments involving local time. We also refer the reader to Crépey [14, section 4.1, Part IV] and Berestycki, Busca, and Florent [4] for results in the case where a is uniformly continuous. Here, we prove the more general case where $a \in \mathcal{M}_Q(\underline{a}, \overline{a})$. This is indeed the case that will be relevant for the study of the calibration problems.

THEOREM 4.3. *Assume $p \in]2, \overline{p}[$. Then the following hold:*

1. *The call price*

$$\overline{Q}^T \ni (t, y) \mapsto \Pi_{T,k}(t, y; a)$$

is the unique $L_{p,loc}(Q^T)$-solution between 0 and S of $BS_{Q^T}(k; a)$. Moreover, it is convex and nondecreasing with respect to S, nondecreasing with the local volatility, and converges to 0 when $S \to 0$, uniformly with t.

2. *The call price*

$$\overline{Q}_{t_0} \ni (T, k) \mapsto \Pi_{T,k}(t_0, y_0; a)$$

is the unique $L_{p,loc}(Q_{t_0})$-solution between 0 and S_0 of $DUP_{Q_{t_0}}(y_0; a)$. Moreover, it is convex and nonincreasing with respect to K, nondecreasing with the local volatility, and converges to 0 when $K \to +\infty$, uniformly with T. Finally, for almost every $t > t_0$, the y-process (3.3) admits a transition probability density between t_0 and t. Discounting this density at rate r, it becomes

$$\gamma_{t_0,y_0}(t, y; a) \equiv e^{-y}(\partial_{y^2}^2 - \partial_y)\Pi_{t,y}(t_0, y_0; a) \ . \tag{4.11}$$

Proof. We proceed by density from the known case of a Lipschitzian function a_n approximating a as in Lemma 3.3. Denoting $(p+\overline{p})/2$ by p', let $\hat{\Pi}$, respectively, Π_n, be the strong solution in $L_{p,loc}(Q^T) \cap L_{p',loc}(Q^T)$ between 0 and S of $BS_{Q^T}(k;\hat{a})$, respectively, $BS_{Q^T}(k;a_n)$.

Since $2 < p < p' < \overline{p} < 3$, it is well known that

$$(\partial^2_{y^2} - \partial_y)\hat{\Pi} \in L_p(Q^T) \cap L_{p'}(Q^T)$$

(see, for instance, Crépey [14, Remark 4.1, Part IV]). Therefore, using Theorem 4.2, there exists an $L_p(Q^T) \cap L_{p'}(Q^T)$-solution $\delta\Pi$ of $BS'_{Q^T}(a;\Gamma)$, where $\Gamma \equiv (a-\hat{a})(\partial^2_{y^2} - \partial_y)\hat{\Pi}$. By linearity, $\Pi \equiv \hat{\Pi} + \delta\Pi$ is then a strong solution in $L_{p,loc}(Q^T) \cap L_{p',loc}(Q^T)$ of $BS_{Q^T}(k;a)$. Moreover,

$$(\partial^2_{y^2} - \partial_y)\Pi \equiv (\partial^2_{y^2} - \partial_y)\hat{\Pi} + (\partial^2_{y^2} - \partial_y)\delta\Pi \in L_p(Q^T) \cap L_{p'}(Q^T)\ .$$

Denote $\Pi_n - \hat{\Pi}$ by $\delta_n\Pi$. By linearity, symmetry, parity, and the results of the theorem in the Lipschitzian case, $\delta_n\Pi$ converges to 0 when $|y| \to +\infty$, uniformly with t, and $\delta_n\Pi$ is a strong solution in $L_{p,loc}(Q^T) \cap L_{p',loc}(Q^T)$ of $BS'_{Q^T}(a_n;\Gamma_n)$, where $\Gamma_n \equiv (a_n - \hat{a})(\partial^2_{y^2} - \partial_y)\hat{\Pi}$. Therefore, by Theorem 4.2, $\delta_n\Pi$ is the strong solution in $L_p(Q^T) \cap L_{p'}(Q^T)$ of $BS'_{Q^T}(a_n;\Gamma_n)$. So $\Pi_n - \Pi = \delta_n\Pi - \delta\Pi$ is the strong solution in $L_p(Q^T) \cap L_{p'}(Q^T)$ of $BS'_{Q^T}(a_n;\Gamma'_n)$, where

$$\Gamma'_n \equiv \Gamma_n - \Gamma + (a_n - a)(\partial^2_{y^2} - \partial_y)\delta\Pi = (a_n - a)(\partial^2_{y^2} - \partial_y)\Pi.$$

Furthermore, Γ'_n converges to 0 in $L_p(Q^T)$ when $n \to +\infty$. Indeed, having fixed $\varepsilon > 0$, let us choose a subset $Q_\varepsilon \equiv Q^T \cap \{|y| \le Y_\varepsilon\}$ such that $\|(\partial^2_{y^2} - \partial_y)\Pi\|_{L_p(Q^c_\varepsilon)} \le \varepsilon$, where $Q^c_\varepsilon \equiv Q^T \setminus Q_\varepsilon$. By Hölder's inequality, it follows, thanks to Lemma 3.3, that

$$\|\Gamma'_n\|^p_{L_p(Q^T)} \le (\|(\partial^2_{y^2} - \partial_y)\Pi\|^p_{L_{p'}(Q^T)} + (\overline{a} - \underline{a})^p)\varepsilon^p$$

for n large enough.

Using estimate (4.3) applied to $\Pi_n - \Pi$, Π then inherits the bounds on Π_n. So $BS^+_{Q^T}(k;a)$ admits an $L_{p,loc}(Q^T)$-solution $\Pi^+ \equiv \Pi$ between 0 and S. Similarly, $BS^-_{Q^T}(k;a)$ admits an $L_{p,loc}(Q^T)$-solution Π^- between 0 and K. We also have symmetric solutions $\Pi^{+/-}_{T,k}$ for $DUP^{+/-}_{Q_{t_0}}(y_0;a)$. Moreover, $\Pi^{+/-} \equiv \Pi^{+/-}_{T,k}$ by passage to the limits in the analogous identities at fixed n. Furthermore, by item 1 of Lemma 3.6, the solutions $\Pi^{+/-}$ and $\Pi^{+/-}_{T,k}$ are the only ones between the required bounds.

The probabilistic representation for Π^- then results from a generalized integrated Itô formula, as in the proof of Theorem 4.2. Since $\Pi^{+/-}$ is the limit of the $\Pi^{+/-}_n$, the probabilistic representation for Π^+ then follows from those for Π^- and $\Pi^{+/-}_n$, using also the call/put parity at a and a_n.

$\Pi^{+/-}$ and $\Pi^{+/-}_{T,k}$ then inherit the monotonicity and convexity properties valid at fixed n by passage to the limit locally uniform over (t,y) and (T,k), respectively. The asymptotic results follow from those, already known, at constant volatility $\underline{a}$ or $\overline{a}$ and from the monotonicity with respect to a.

Finally, by standard arguments developed, for instance, in Stroock and Varadhan [37, proof of Theorem 9.1.9, p. 224], estimate (4.3), or merely (4.6), valid for all $\Gamma \in L_p(Q^T)$, enforces the existence of a transition probability density between t_0 and t for the process y for almost every $t > t_0$.

Then, by general arguments set out, for instance, in Crépey [14, section 4.1, Part IV], independent of the Lipschitzian assumption on a therein, the discounted density for the process S is $\partial^2_{S^2}\Pi_{t,S}(t_0, S_0; a)$, whence, after a change of variables, we obtain the expression for γ. ∎

The following proposition gathers a few consequences of the previous results that will be useful in the following study of the calibration problems. The proposition is stated for $\Pi \equiv \Pi^+$. The analogous statements for $\Pi \equiv \Pi^-$ follow by parity. We then also have the symmetric statements in the variables (T, k). Recall that h and h' denote elements of $H^1(Q)$.

PROPOSITION 4.4. *Assume $p \in]2, \overline{p}[$.*

1. *Then*

$$\|(\partial^2_{y^2} - \partial_y)\Pi\|_{L_p(Q^T)} \;\leq\; C_p \;, \tag{4.12}$$

where $C_p \equiv C_p(\underline{t}, \overline{T}, \overline{k}; R, \underline{a}, \overline{a})$.

2. *The price Π is locally θ-Hölderian, jointly with respect to $(t_0, y_0), (T, k)$, uniformly with $q, r \in [0, R]$, $a \in \mathcal{M}_Q(\underline{a}, \overline{a})$.*

3. *Further define $p' = (2 + p)/2$, $p'' = (2 + p')/2$, and $\Gamma \equiv h(\partial^2_{y^2} - \partial_y)\Pi$. Then*

$$\|\Gamma\|_{L_{p'}(Q^T)} \;\leq\; C'_{p'}\; \|h\|_{H^1(Q)} \;,$$

where $C'_{p'} \equiv C'_{p'}(\underline{t}, \overline{T}, \overline{k}; R, \underline{a}, \overline{a})$. Then let $d\Pi$, or $d\Pi_{T,k}(\cdot; a).h$, be the $L_{p'}(Q^T)$-solution of $BS'_{Q^T}(a;\, \Gamma)$. Furthermore, let Γ' and $d\Pi'$ be defined as Γ and $d\Pi$ with h' instead of h, and

$$d\Gamma \;\equiv\; h'(\partial^2_{y^2} - \partial_y)d\Pi \;+\; h(\partial^2_{y^2} - \partial_y)d\Pi' \;.$$

Then

$$\|d\Gamma\|_{L_{p''}(Q^T)} \;\leq\; C''_{p''}\; \|h\|_{H^1(Q)}\; \|h'\|_{H^1(Q)} \;,$$

where $C''_{p''} \equiv C''_{p''}(\underline{t}, \overline{T}, \overline{k}; R, \underline{a}, \overline{a})$. We shall then denote by $d^2\Pi$, or $d^2\Pi_{T,k}(\cdot; a).(h, h')$, the $L_{p''}(Q^T)$-solution of $BS'_{Q^T}(a; d\Gamma)$.

4. *We have*

$$\begin{aligned}
\|d\Pi\|_{\mathcal{C}^0_\theta(\overline{Q}^T)} &\leq C'\; \|d\Pi\|_{W^{1,2}_p(Q^T)} \leq\; C'C\; \|h\|_{H^1(Q)},\\
\|d^2\Pi\|_{\mathcal{C}^0_\theta(\overline{Q}^T)} &\leq C'\; \|d^2\Pi\|_{W^{1,2}_p(Q^T)} \leq C'C\; \|h\|_{H^1(Q)}\|h'\|_{H^1(Q)} \;,
\end{aligned}$$

where $C' \equiv C'_p$ is as in (3.4), and $C \equiv C_p(\underline{t}, \overline{T}, \overline{k}; R, \underline{a}, \overline{a})$. Moreover, if $a + h \in \mathcal{M}_Q(\underline{a}, \overline{a})$, let us define, for $\varepsilon \in]0, 1[$,

$$\begin{aligned}
\varepsilon^{-1}\delta_\varepsilon\Pi &\equiv \varepsilon^{-1}[\Pi_{T,k}(\cdot; a + \varepsilon h) - \Pi_{T,k}(\cdot; a)],\\
\varepsilon^{-1}\delta_\varepsilon d\Pi &\equiv \varepsilon^{-1}[d\Pi_{T,k}(\cdot; a + \varepsilon h).h' - d\Pi_{T,k}(\cdot; a).h'] \;.
\end{aligned}$$

When $\varepsilon \to 0$, $\varepsilon^{-1}\delta_\varepsilon\Pi$ and $\varepsilon^{-1}\delta_\varepsilon d\Pi$ converge in $\mathcal{C}^0_\theta(\overline{Q}^T) \cap W^{1,2}_p(Q^T)$, respectively, to $d\Pi$ and $d^2\Pi$.

5. *Assume furthermore that a and, for $n \in \mathbb{N}^\star$, a_n, belong to $a_0 + H^1_Q(\underline{a}, \overline{a})$, where $a_n - a$ converges to 0 weakly in $H^1(Q)$ when $n \to +\infty$. Then $\Pi_n \equiv \Pi_{T,k}(\cdot; a_n)$ converges to $\Pi \equiv \Pi_{T,k}(\cdot; a)$ in $\mathcal{C}^0_\theta(\overline{Q}^T) \cap W^{1,2}_p(Q^T)$.*

Notice that $d\Pi$ and $d^2\Pi$ in this proposition are well defined by Theorem 4.2.

Proof. The proof is deferred to Appendix B. ∎

5. Stability.

5.1. The ill-posed calibration problems. Let us now give a rigorous statement of the calibration problems. From now on, we assume $p \in]2, \overline{p}[$, and we will denote by $\overset{\circ}{W}{}_p^{1,2}(Q_{t_0})$ the set of functions in $W_p^{1,2}(Q_{t_0})$ that vanish at time t_0. We also fix a finite subset $\mathcal{F} \subset Q_{t_0}$ with $|\mathcal{F}| = M \in \mathbb{N}^\star$. Then we define the following nonlinear *pricing functional*:

$$a_0 + H_Q^1(\underline{a}, \overline{a}) \ni a \overset{\Pi}{\longmapsto} \Pi(a) \in \Pi_0 + \overset{\circ}{W}{}_p^{1,2}(Q_{t_0}) ,$$

where Π_0, respectively, $\Pi(a)$, denotes the $L_{p,loc}(Q_{t_0})$-solution between 0 and S_0 of $DUP_{Q_{t_0}}(y_0; a_0)$, respectively, $DUP_{Q_{t_0}}(y_0; a)$. Recall that $a_0 \in \mathcal{M}_Q(\underline{a}, \overline{a})$ denotes the *prior* of the calibration problem (see section 2.3).

PROPOSITION 5.1. *The pricing functional* Π *and the restriction* $\Pi_{|\mathcal{F}}$ *are well defined on the closed convex subset* $a_0 + H_Q^1(\underline{a}, \overline{a})$ *of* $a_0 + H^1(Q)$. *Moreover, we have the following:*

1. *(Compactness.)* Π *and* $\Pi_{|\mathcal{F}}$ *map weakly convergent sequences into strongly convergent ones.*

2. *(Differentiability.)* Π *and* $\Pi_{|\mathcal{F}}$ *are twice Gâteaux differentiable.*

3. *(Perturbations of the operator.)* $\Pi_{|\mathcal{F}}$ *has* θ*-Hölderian dependence with respect to* (t_0, y_0) *and* $\mathcal{F}$.

Proof. By Theorems 4.2 and 4.3, Π and $\Pi_{|\mathcal{F}}$ are well defined. Now, points 1, 2, and 3, respectively, follow from the results symmetric to Proposition 4.4(5), 4.4(4), and 4.4(2) in the variables (T, k). □

DEFINITION 5.2. *By* the continuous calibration problem *with data*

$$\widetilde{\Pi} \in \Pi_0 + \overset{\circ}{W}{}_p^{1,2}(Q_{t_0}) ,$$

respectively, the discrete calibration problem *with data* $\pi \in \mathbb{R}^M$, *we will mean, finding an* $a \in a_0 + H_Q^1(\underline{a}, \overline{a})$ *such that*

$$\widetilde{\Pi}_{T,k} = \Pi_{T,k}(t_0, y_0; a) , \quad (T, k) \in Q_{t_0},$$

respectively,

$$\pi_{T,k} = \Pi_{T,k}(t_0, y_0; a) , \quad (T, k) \in \mathcal{F} .$$

Data for which this is possible will be said to be calibrateable.

Remark 5.3. To fix notation, we thus consider the calibration problems with European call option prices. However, by symmetry and parity, all the results below extend straightaway to the following situations:

1. (Continuous problem.) Calibration from European put option prices.
2. (Discrete problem.) Calibration from European call and put option prices.

A nonlinear inverse problem is said to be *ill-posed* at any data set around which the direct operator (here, the pricing functional Π or $\Pi_{|\mathcal{F}}$) is not continuously invertible.

THEOREM 5.4. *For every continuous function* $a \in a_0 + H_Q^1(\underline{a}, \overline{a})$, *the continuous calibration problem is ill-posed at* $\widetilde{\Pi} \equiv \Pi(a)$, *and the discrete calibration problem is ill-posed at* $\pi \equiv \Pi_{|\mathcal{F}}(a)$.

Proof. See Appendix C for the proof. □

5.2. Stabilization by Tikhonov regularization. The best-known stabilization method for ill-posed nonlinear inverse problems is Tikhonov regularization [38, 21], which we now consider. The properties of the nonlinear pricing functional Π, summed up in Proposition 5.1, will allow us to apply the general theory surveyed, for instance, in Engl, Hanke, and Neubauer [21, Chapter 10].

In practice, market prices π are defined as bid-ask spreads. Moreover, $\widetilde{\Pi}$ depends on an interpolation procedure. Therefore, the actual set of observed prices, or input data, for the calibration, π^δ or $\widetilde{\Pi}^\delta$, is only known up to some noise δ. Moreover, any numerical procedure used to tackle the discrete calibration problem entails some computational burden η. Furthermore, the local volatility function is calibrated at the current phase (t_0, y_0) and set $\mathcal{F}$, and used later at the perturbed phase (t_0^μ, y_0^μ) and set $\mathcal{F}_\mu$. The Tikhonov regularization method allows one to overcome such data noise, computational burden, and perturbations of the operator.

DEFINITION 5.5. (Continuous problem.) *By an* α-solution *of the continuous calibration problem with prior* a_0 *and noisy data*

$$\widetilde{\Pi}^\delta \in \Pi_0 + \overset{\circ}{W}{}_p^{1,2}(Q_{t_0}) \,,$$

we will mean, in $a_0 + H_Q^1(\underline{a}, \overline{a})$*, any* a_α^δ *such that for every* a*,*

$$J_\alpha^\delta \left(a_\alpha^\delta\right) \leq J_\alpha^\delta \left(a\right),$$

where

$$2J_\alpha^\delta (a) \equiv \left\| \Pi \left(t_0, y_0, a\right) - \widetilde{\Pi}^\delta \right\|^2_{W_p^{1,2}(Q_{t_0})} + \alpha \left\| a - a_0 \right\|^2_{H^1(Q)} \,.$$

(Discrete problem.) *By an* α-solution *of the discrete calibration problem with prior* a_0*, noisy data* $\pi^\delta \in \mathbb{R}^M$*, perturbed parameters* $(t_0^\mu, y_0^\mu) \in Q$*,* $\mathcal{F}_\mu \subset Q_{t_0^\mu}$ *(*$|\mathcal{F}_\mu| = M$*), and computational burden* $\eta \geq 0$*, we will mean, in* $a_0 + H_Q^1(\underline{a}, \overline{a})$*, any* $a_\alpha^{\delta,\mu,\eta}$ *such that for every* a*,*

$$J_\alpha^{\delta,\mu} \left(a_\alpha^{\delta,\mu,\eta}\right) \leq J_\alpha^{\delta,\mu} \left(a\right) + \eta,$$

where

$$2J_\alpha^{\delta,\mu} (a) \equiv \left\| \Pi_{|\mathcal{F}_\mu} \left(t_0^\mu, y_0^\mu, a\right) - \pi^\delta \right\|^2_{\mathbb{R}^M} + \alpha \left\| a - a_0 \right\|^2_{H^1(Q)} \,.$$

Such α-solutions do exist because of Proposition 5.1(1). We shall not address in this paper the problem of the uniqueness of the unregularized calibration problems, or of the regularized problems for arbitrary values of the regularization parameter α. However, at least for the discrete problem, one has the following result when α tends to $+\infty$. The intuition behind this result is that when α tends to $+\infty$, the regularization term becomes dominant and enforces the convexity of the cost criterion as a whole.

THEOREM 5.6. *There exists* $C \equiv (1 + \overline{\pi}^\delta) MC_p(\underline{t}, \overline{y}_0, \overline{T}; R, \underline{a}, \overline{a})$ *such that the cost criterion* $J \equiv J_\alpha^{\delta,\mu}$ *is* C*-strongly convex on* $a_0 + H_Q^1(\underline{a}, \overline{a})$ *for every* $\alpha \geq 2C$*. Here,* $\overline{y}_0$ *and* $\overline{\pi}^\delta$ *denote bounds on* $|y_0^\mu|$ *and* $\pi_{T,k}^\delta$ *for* $(T, k) \in \mathcal{F}_\mu$*.*

$J_\alpha^{\delta,\mu}$ then admits a unique minimum, which depends continuously upon (t_0^μ, y_0^μ), $\mathcal{F}_\mu$, and π^δ. Otherwise said, the minimization problem of $J_\alpha^{\delta,\mu}$ is well-posed in the sense of Hadamard.

Proof. By the chain rule, we have

$$\begin{aligned} d^2J(a).(h,h') \;\equiv\; & \alpha\langle h,h'\rangle_{H^1(Q)} \\ & + \sum_{(T,k)\in\mathcal{F}_\mu} d\Pi_{T,k}(t_0^\mu,y_0^\mu;a).h \; d\Pi_{T,k}(t_0^\mu,y_0^\mu;a).h' \\ & + \sum_{(T,k)\in\mathcal{F}_\mu} \left(\Pi_{T,k}(t_0^\mu,y_0^\mu;a)-\pi^\delta_{T,k}\right)\; d^2\Pi_{T,k}(t_0^\mu,y_0^\mu;a).(h,h') \;. \end{aligned}$$

For $a,b\in a_0+H^1_Q(\underline{a},\overline{a})$ and $\varepsilon\in\,]0,1[$, let us define $a_\varepsilon\equiv(1-\varepsilon)a+\varepsilon b$, $J_\varepsilon\equiv J(a_\varepsilon)$. Using Proposition 5.1(2) and the bound $e^{y_0^\mu}$ on $|\Pi|$, it follows, denoting by $'$ the derivative with respect to ε, that

$$\begin{aligned} \langle\nabla J(b)-\nabla J(a),b-a\rangle_{H^1(Q)} = J_1'-J_0' \;=\; & \int_0^1 J_\varepsilon''d\varepsilon \\ = \int_0^1 d^2J(a_\varepsilon).(b-a,b-a) \geq (\alpha-(1+e^{y_0^\mu}+\overline{\pi}^\delta)MC)\;\|b-a\|^2_{H^1(Q)}\;, & \end{aligned}$$

where $C\equiv C_p(\underline{t},\overline{y}_0,\overline{T};R,\underline{a},\overline{a})$. ∎

Moreover, Tikhonov regularized solutions of the calibration problems at arbitrary level $\alpha>0$ are *stable* in the following meaning.

THEOREM 5.7. (Stability, continuous problem.) *Assume* $\widetilde{\Pi}^{\delta_n}\to\widetilde{\Pi}^\delta$ *when* $n\to+\infty$. *Then any sequence of* α*-solutions* $a_\alpha^{\delta_n}$ *admits a subsequence which converges towards an* α*-solution* a_α^δ.

(Stability, discrete problem.) *Assume*

$$\pi^{\delta_n},\quad (t_0^{\mu_n},y_0^{\mu_n}),\quad \mathcal{F}_{\mu_n},\quad \eta_n\longrightarrow\pi^\delta,\quad (t_0^\mu,y_0^\mu),\quad \mathcal{F}_\mu,\quad \eta\equiv 0,$$

when $n\to+\infty$. *Then any sequence of* α*-solutions* $a_\alpha^{\delta_n,\mu_n,\eta_n}$ *admits a subsequence which converges towards an* α*-solution* $a_\alpha^{\delta,\mu,\eta\equiv0}$.

Notice that this convergence is strong in $H^1(Q)$.

Proof. Using Proposition 5.1(1), this results directly from Theorem 2.1 in Engl, Kunisch, and Neubauer [22], supplemented by Remark 3.4 in Binder et al. [5], for the continuous problem. For the discrete problem, the proof is an immediate adaptation of the one in [22, Theorem 2.1], using items 1 and 3 of Proposition 5.1. ∎

6. Convergence and convergence rates.

6.1. Convergence. We are going to see that the Tikhonov regularization method behaves as an approximating scheme for the pseudoinverse of Π or $\Pi_{|\mathcal{F}}$. By *pseudoinverse*, we mean the operator that maps calibrateable data $\widetilde{\Pi}$ or π to an element a which minimizes $\|a-a_0\|$ over the set of all preimages of $\widetilde{\Pi}$ or π through Π or $\Pi_{|\mathcal{F}}$.

DEFINITION 6.1 (a_0-MNS). *Given calibrateable data, we shall call an* a_0-minimum norm solution *(a_0-MNS) of the calibration problem any solution a that minimizes $\|a-a_0\|$ over the set of all solutions.*

Such an a_0-MNS a exists for all calibrateable data. But it may be nonunique, since the pricing functional Π is nonlinear.

THEOREM 6.2. (Convergence, continuous problem.) *Given calibrateable data* $\widetilde{\Pi}$, *suppose that*

$$\begin{aligned} \left\|\widetilde{\Pi}-\widetilde{\Pi}^{\delta_n}\right\|_{W_p^{1,2}(Q_{t_0})} \leq \delta_n \quad & \textit{for} \quad n\in\mathbb{N}, \\ \alpha_n,\;\delta_n^2/\alpha_n \quad \longrightarrow 0 \quad & \textit{when} \quad \to+\infty. \end{aligned}$$

Then any sequence $a^{\delta_n}_{\alpha_n}$ admits a subsequence which converges towards an a_0-MNS a. Moreover, $a^{\delta_n}_{\alpha_n} \to a$ if a is the unique a_0-MNS of the calibration problem at $\widetilde{\Pi}$.

(Convergence, discrete problem.) *Given calibrateable data π, suppose that*

$$\left\|\pi - \pi^{\delta_n}\right\|_{\mathbb{R}^M} \leq \delta_n, \quad |t_0 - t_0^{\mu_n}| \vee |y_0 - y_0^{\mu_n}| \vee \|\mathcal{F} - \mathcal{F}_{\mu_n}\| \leq \mu_n \ \text{ for } \ n \in \mathbb{N},$$
$$\alpha_n, \quad \delta_n^2/\alpha_n, \quad \mu_n^{2\theta}/\alpha_n, \quad \eta_n/\alpha_n \longrightarrow 0 \ \text{ when } \ n \to +\infty.$$

Then any sequence $a^{\delta_n,\mu_n,\eta_n}_{\alpha_n}$ admits a subsequence which converges towards an a_0-MNS a. Moreover, $a^{\delta_n,\mu_n,\eta_n}_{\alpha_n} \to a$ if a is the unique a_0-MNS of the calibration problem at π.

Proof. Using Proposition 5.1(1), this follows directly from Theorem 2.3 in Engl, Kunisch, and Neubauer [22], supplemented by Remark 3.4 in Binder et al. [5], for the continuous problem. For the discrete problem, it results, for instance, from Kunisch and Geymayer [27, Proposition 1], using items 1 and 3 of Proposition 5.1. □

Following Engl, Hanke, and Neubauer [21, Proposition 3.11 and Remark 3.12], there can be, for the convergence of such regularized schemes towards solutions of ill-posed inverse problems, no uniform rate over all calibrateable data. In fact, this presents a generic character for any method of resolution, Tikhonov or otherwise. It is therefore important to be able to specialize subsets of $a_0 + H^1_Q(\underline{a}, \overline{a})$ on which such uniform rates may be exhibited.

6.2. Convergence rates. We first have the following abstract statement. Let $d\Pi_{|\mathcal{F}}(a)^\star$ and $d\Pi(a)^\star$ denote the adjoints of the operators $d\Pi_{|\mathcal{F}}(a)$ and $d\Pi(a)$, respectively. That is to say, by definition,

$$\langle h\ , d\Pi_{|\mathcal{F}}(a)^\star \lambda \rangle_{H^1(Q)} \ = \sum_{(T,k)\in\mathcal{F}} \lambda_{T,k} d\Pi_{T,k}(a).h\ ; \quad (h,\lambda) \in H^1(Q) \times \mathbb{R}^M,$$

respectively,

$$\langle h\ , d\Pi(a)^\star \lambda \rangle_{H^1(Q)} \ = \ \langle d\Pi(a).h\ , \lambda \rangle_{W^{1,2}_p(Q_{t_0}), W^{1,2}_\rho(Q_{t_0})}\ ; \quad (h,\lambda) \in H^1(Q) \times W^{1,2}_\rho(Q_{t_0}),$$

where $p^{-1} + \rho^{-1} = 1$, and where the last bracket denotes the *duality bracket* between λ and $d\Pi(a).h$.

THEOREM 6.3. (Convergence rates, continuous problem.) *There exists $C_p \equiv C_p(\underline{t}, \overline{y}_0, \overline{T}; R, \underline{a}, \overline{a})$ such that for every a_0-MNS a of the calibration problem at $\widetilde{\Pi}$ with*

$$a - a_0 = d\Pi(a)^\star \lambda \tag{6.1}$$

for some $\|\lambda\|_{W^{1,2}_\rho(Q_{t_0})} \leq C_p$, then

$$\|a^\delta_\alpha - a\|_{H^1(Q)} = \mathrm{O}(\delta^{\frac{1}{2}}),$$

whenever

$$\left\|\widetilde{\Pi} - \widetilde{\Pi}^\delta\right\|_{W^{1,2}_p(Q_{t_0})} \leq \delta, \quad \alpha \ \sim \ \delta\ .$$

(Convergence rates, discrete problem.) *There exists $C_p \equiv C_p(\underline{t}, \overline{y}_0, \overline{T}; R, \underline{a}, \overline{a})$ such that for every a_0-MNS a of the calibration problem at π with*

$$a - a_0 = d\Pi_{|\mathcal{F}}(t_0, y_0; a)^\star \lambda \tag{6.2}$$

for some $\|\lambda\|_{\mathbb{R}^M} \leq C_p/\sqrt{M}$*, then*

$$\|a_\alpha^{\delta,\mu,\eta} - a\|_{H^1(Q)} = \mathrm{O}(\delta^{\frac{1}{2}} + \mu^{\frac{\theta}{2}}),$$

whenever

$$\left\|\pi - \pi^\delta\right\|_{\mathbb{R}^M} \leq \delta\,, \quad |t_0 - t_0^\mu| \vee |y_0 - y_0^\mu| \vee \|\mathcal{F} - \mathcal{F}_\mu\| \leq \mu\,, \quad \alpha \;\sim\; \delta \vee \mu^\theta\,, \quad \eta = \mathrm{O}\left(\delta^2\right).$$

Therefore a is the only a_0-MNS satisfying condition (6.1) or (6.2).

Proof. (Continuous problem.) Using items 1 and 2 of Proposition 5.1, this follows from Engl, Hanke, and Neubauer [21, Theorem 10.4 and Remark 10.5] by noticing that the proof therein readily extends from their Hilbert → Hilbert to our Hilbert → reflexive Banach setting, by reading duality brackets instead of inner products.

(Discrete problem.) Using Proposition 5.1, this follows from Kunisch and Geymayer [27, Theorem 2 and Remark iv, p. 86]. □

Remark 6.4. Kunisch and Geymayer [27, Theorem 2] assume that a belongs to the interior of $a_0 + H_Q^1(\underline{a}, \overline{a})$. However, this cannot be realized in our case. Indeed, $a_0 + H_Q^1(\underline{a}, \overline{a})$ has an empty interior. But this assumption is not used as long as discretization of the source space is not dealt with.

Except in the trivial case where $a \equiv a_0$, conditions (6.1)–(6.2) may seem rather abstract. Whether there is a neighborhood around a_0 such that they are satisfied is an open question. However, in the case where a is uniformly continuous with respect to its space variable y, one can derive a more explicit formulation of (6.2). In the following, let $\widetilde{\nabla}\Pi_{T,k}$, not to be mistaken with the Gâteaux derivative of Π in $H^1(Q)$, denote the following function on Q, parameterized by (t_0, y_0, T, k) and a:

$$\widetilde{\nabla}\Pi_{T,k}(t,y) \;\equiv\; \mathbf{1}_{\{t_0<t<T\}} e^{-y}(\partial_{y^2}^2 - \partial_y)\Pi_{t,y}(t_0, y_0; a)(\partial_{y^2}^2 - \partial_y)\Pi_{T,k}(t, y; a)\,.$$

Lemma 6.5. *For* $(T,k) \in \mathcal{F}$*,*

$$d\Pi_{T,k}(t_0, y_0; a)\,.\,h \;=\; \int\!\!\int_Q \widetilde{\nabla}\Pi_{T,k}\; h\,.$$

Proof. Indeed, this is just the probabilistic representation (4.2) for $d\Pi$, given the expression for γ in Theorem 4.3(2) and the L_p estimate on Γ in Proposition 4.4(3). □

Theorem 6.6. *Let* $a \in a_0 + H_Q^1(\underline{a}, \overline{a})$ *be uniformly continuous with respect to its space variable* y*. Then the following hold:*

1. $\widetilde{\nabla}\Pi_{T,k} \in L_2(Q)$ *for* $(T,k) \in \mathcal{F}$*.*
2. $\Lambda \equiv d\Pi_{|\mathcal{F}}(a)^\star\lambda$ *is the unique solution in* $H^2(Q)$ *of the following problem:*

$$\left\{\begin{array}{l} \Lambda - \Delta\Lambda = \sum_{(T,k)\in\mathcal{F}} \lambda_{T,k}\widetilde{\nabla}\Pi_{T,k}\,, \quad Q\text{-a.e.},\\ \partial_n\Lambda \;=\; 0\,, \quad \partial Q\text{-a.e.} \end{array}\right. \tag{6.3}$$

3. *Condition* (6.2) *means that* (6.3) *holds with* $\Lambda \equiv a - a_0$ *for some*

$$\|\lambda\|_{\mathbb{R}^M} \;\leq\; C_p(\underline{t}, \overline{y}_0, \overline{T}; R, \underline{a}, \overline{a})/\sqrt{M}.$$

Notice that by Theorem 3.2(3), the normal derivative $\partial_n\Lambda \in L_2(\partial Q)$ is well defined for $\Lambda \in H^2(Q)$.

Proof. 1. According to Proposition 4.4(3),

$$(\partial_{y^2}^2 - \partial_y)\Pi_{T,k}(t,y;a) \in L_p\left(\left]\frac{t_0+T}{2}, T\right[\times \mathbb{R}\right).$$

On the other hand, we have by Stroock and Varadhan [37, Theorem 9.1.9, equation (1.35)],

$$e^{-y}(\partial_{y^2}^2 - \partial_y)\Pi_{t,y}(t_0,y_0;a) = \gamma_{t_0,y_0}(t,y;a) \in L_q\left(\left]\frac{t_0+T}{2}, T\right[\times \mathbb{R}\right)$$

for every $1 \leq q < +\infty$. More precisely,

$$\|\gamma_{t_0,y_0}(\cdot;a)\|_{L_q(]\frac{t_0+T}{2},T[\times\mathbb{R})} \leq C_q^\omega(\underline{t}, \overline{T}, R, \underline{a}, \overline{a}) \,,$$

where ω denotes a modulus of continuity of a with respect to y. Hence $\widetilde{\nabla}\Pi_{T,k} \in L_2(]\frac{t_0+T}{2}, T[\times\mathbb{R})$ by Hölder's inequality. By symmetry and parity, we can conclude that $\widetilde{\nabla}\Pi_{T,k} \in L_2(Q)$.

2. Therefore, using Lemma 6.5, the adjunction relations for Λ can be written as

$$\langle \Lambda\ , h\rangle_{H^1(Q)} \ = \sum_{(T,k)\in\mathcal{F}} \lambda_{T,k}\ \langle \widetilde{\nabla}\Pi_{T,k}, h\rangle_{L_2(Q)}\ , \quad h \in H^1(Q)\ . \tag{6.4}$$

It is then known that the adjoint $\Lambda \in H^1(Q)$ belongs in fact to $H^2(Q)$—see, for instance, Bensoussan and J.-L. Lions [3, Theorem 5.10, Chapter 2, and the footnote on p. 96]. We can then apply the generalized Green formula to identity (6.4) and conclude in a classic way, using Theorem 3.2(4); see, for example, Larrouturou and P. L. Lions [30, p. 150, step 6, Interpretation of the variational formulation].

3. Item 3 follows immediately from 2. □

Remark 6.7.

1. The condition in Theorem 6.6(3), which ensures a convergence rate in $\mathrm{O}(\delta^{\frac{1}{2}} + \mu^{\frac{\theta}{2}})$, is very severe, since its implies that $(\mathrm{Id} - \Delta).(a - a_0)$ belongs to the $\leq M$-dimensional subspace of $L_2(Q)$ spanned by the $\widetilde{\nabla}\Pi_{T,k}$, $(T,k) \in \mathcal{F}$ for sufficiently small coefficients $\lambda_{T,k}$.

2. This condition is both a closedness and smoothness condition of a with respect to a_0, which says that, as already noted elsewhere, "Tikhonov regularization can only resolve smooth details fast" [36, p. 611]. Indeed, one then has the following $H^2(Q)$ estimate from regularity theory for elliptic equations (method of tangential translations; see, for instance, Brézis [9, pp. 181 and 184]):

$$\|a - a_0\|_{H^2(Q)} \ \leq\ \sqrt{M}\ C_p^\omega(\underline{t}, \overline{y}_0, \overline{T}; R, \underline{a}, \overline{a})\ \|\lambda\|_{\mathbb{R}^M},$$

where ω denotes a modulus of continuity of a with respect to y.

3. At least in the Hilbert → Hilbert setting of the discrete problem, there exist conditions stronger than (6.2) ensuring better convergence rates, typically in $\mathrm{O}(\delta^{\frac{2}{3}})$; see, for instance, [32, 33, 21]. But these conditions require that a be interior to the domain of definition of the direct operator—see, for instance, Neubauer [32, equation (2.5)]. As already observed above, this cannot be realized in our case. Indeed, $H_Q^1(\underline{a}, \overline{a})$ has an empty interior. Nonetheless, the reader is referred to Neubauer and Scherzer [33, section 3] for a special case in which an $\mathrm{O}(\delta^{\frac{2}{3}})$ convergence rate is proved, although the domain of definition of the direct operator has empty interior.

7. Conclusion. Having established $W_p^{1,2}$ estimates for Black–Scholes and Dupire equations with measurable ingredients, we have shown that the problem of inverting observed vanilla option prices into a local volatility function, in a generalized Black–Scholes model, fits into the frame of the Tikhonov regularization method. Moreover, this holds true both when the option prices form a continuum and when they consist of a finite set. We were then able to derive results for stability, convergence, and convergence rates for this method. Discretization and effective implementation, as well as numerical results, can be found in [16]. This work also deals with an extension of the numerical implementation to the problem of calibration from American option prices. With respect to this, an open problem is whether the theoretical results obtained in the present article relating to calibration from European option prices, in a generalized Black–Scholes model, may be extended to calibration from American option prices. Another more incidental open problem is whether the continuity assumption is necessary in Theorem 6.6.

Appendix A. A technical lemma. The following lemma justifies the passage to the limit at the end of the proof of Theorem 4.2. Although it is an adaptation of Theorem 3.8 in Caffarelli et al. [10], using also Theorem 2.8 in Crandall, Kocan, and Swiech [13], we give the proof in detail for completeness. The notation is the same as above.

LEMMA A.1. *Let us be given $\Gamma \in \mathcal{D}(\overline{Q}^T)$, and $2 < p' < p$. For $n \in \mathbb{N}$, let φ_n be an $L_{p',loc}(Q^T)$-solution of $BS'_{Q^T}(a_n;\Gamma)$, where a_n is a Lipschitzian approximation of a as in Lemma 3.3. Assume the existence of a function $\varphi \in W^{1,2}_{p,loc}(Q^T)$ such that $\varphi_n \to \varphi$ when $n \to +\infty$, locally uniformly on $\overline{Q}^T$. Then φ is an $L_{p,loc}(Q^T)$-solution of $BS'_{Q^T}(a;\Gamma)$.*

Proof. The proof proceeds by contradiction. Assume that φ is, say, no $L_{p,loc}(Q^T)$-viscosity subsolution of $BS'_{Q^T}(a;\Gamma)$. Therefore, there exist open nonempty bounded intervals I and J, a rectangle $Q' = I \times J \subseteq Q^T$ centered at a point $(t_0, y_0) \in Q^T$, and a test function $\psi \in W_p^{1,2}(Q')$ such that

$$-\partial_t\psi - (r - q - a(t,y))\,\partial_y\psi - a(t,y)\partial^2_{y^2}\psi + r\varphi \;>\; \Gamma + \varepsilon \quad \text{on} \quad Q', \tag{A.1}$$

$$(\varphi - \psi)(t_0, y_0) = 0,\ \varphi - \psi < -\delta \quad \text{on} \quad \partial_p Q'\ . \tag{A.2}$$

Moreover, due to the Hölderian character of φ and ψ through the Sobolev embedding (3.1) on Q', one can assume

$$\varphi - \psi < -\frac{\delta}{2} \quad \text{on } \partial_p Q'' \tag{A.3}$$

for some subrectangle Q'' with the same properties as Q' and $\overline{Q}'' \subset Q'$.

We are going to construct a sequence of functions ψ_n (hence, $\psi+\psi_n) \in W^{1,2}_{p',loc}(Q')$ such that

$$\psi_n \;\to\; 0 \quad \text{in } L_\infty(Q'') \quad \text{as } n \to \infty \tag{A.4}$$

and for n large enough

$$\begin{aligned} &-\partial_t(\psi+\psi_n) - (r - q - a_n(t,y))\,\partial_y(\psi+\psi_n) - a_n(t,y)\partial^2_{y^2}(\psi+\psi_n) + r\varphi_n \\ &\qquad \geq \Gamma + \varepsilon \text{ on } Q''. \end{aligned} \tag{A.5}$$

Then by (A.2), (A.3), (A.4), and the assumed local uniform convergence of φ_n to φ, $\varphi_n - (\psi+\psi_n)$ will be larger at (t_0, y_0) than anywhere else on $\partial_p Q''$ for n large enough.

In view of (A.5), this contradicts the assumption that φ_n is an $L_{p',loc}(Q^T)$-viscosity solution of $BS'_{Q^T}(a_n;\Gamma)$.

To construct ψ_n, notice that by (A.1), we have on Q', for ψ_n arbitrary in $W^{1,2}_{p',loc}(Q')$,

$$\begin{aligned}
&-\partial_t(\psi+\psi_n)-(r-q-a_n(t,y))\,\partial_y(\psi+\psi_n)-a_n(t,y)\partial^2_{y^2}(\psi+\psi_n)+r\varphi_n-\Gamma\\
&\geq \varepsilon+(a-a_n)(\partial^2_{y^2}-\partial_y)\psi-r(\varphi-\varphi_n)\\
&\quad-\partial_t\psi_n-(r-q-a_n(t,y))\,\partial_y\psi_n-a_n(t,y)\partial^2_{y^2}\psi_n\\
&\geq \varepsilon+\Gamma_n-\partial_t\psi_n-(R+\overline{a})\,|\partial_y\psi_n|-\overline{a}\big(\partial^2_{y^2}\psi_n\big)^+-\underline{a}\big(\partial^2_{y^2}\psi_n\big)^-,
\end{aligned}$$

where

$$\Gamma_n\equiv(a-a_n)(\partial^2_{y^2}-\partial_y)\psi-r(\varphi-\varphi_n)\;\to\;0\quad\text{in } L_{p'}(Q')\quad\text{as } n\to\infty.$$

Now, choose ψ_n to be, by Theorem 2.8 in Crandall, Kocan, and Swiech [13], the $L_{p',loc}(Q')$-solution of the following problem:

$$\begin{cases}\partial_t\psi_n+(R+\overline{a})\,|\partial_y\psi_n|+\overline{a}\,\big(\partial^2_{y^2}\psi_n\big)^+-\underline{a}\,\big(\partial^2_{y^2}\psi_n\big)^- \;=\;\Gamma_n\quad\text{on } Q',\\ \psi_n\;=\;0\quad\text{on } \partial_pQ',\end{cases}$$

with estimate

$$\|\psi_n\|_{W^{1,2}_{p'}(Q'')}\;\leq\;C\;\|\Gamma_n\|_{L_{p'}(Q')},$$

$C\equiv C_{p'}(R,\underline{a},\overline{a},Q',Q'')$ independent of n. Considering the Sobolev embedding (3.1) on Q'', this furnishes the desired sequence ψ_n. ∎

Appendix B. Proof of Proposition 4.4.

1. By Theorem 4.3(1), let us consider Π, respectively, $\hat{\Pi}$, the $L_{p,loc}(Q^T)$-solution between 0 and S of $BS_{Q^T}(k;a)$, respectively, $BS_{Q^T}(k;\hat{a})$. Then by linearity, symmetry, parity, and the asymptotic results in Theorem 4.3(1), $\delta\Pi\equiv\Pi-\hat{\Pi}$ converges to 0 when $|y|\to+\infty$, uniformly with t, and $\delta\Pi$ is an $L_{p,loc}(Q^T)$-solution of $BS'_{Q^T}(a;\Gamma)$, where $\Gamma\equiv(a-\hat{a})(\partial^2_{y^2}-\partial_y)\hat{\Pi}$. Now, it is well known that

$$\|(\partial^2_{y^2}-\partial_y)\hat{\Pi}\|_{L_p(Q^T)}\;\leq\;C_p(\underline{t},\overline{T},\overline{k},R,\underline{a},\overline{a})$$

(see, for instance, Crépey [14, Remark 4.1, Part IV]). Using also (4.3), this gives (4.12).

2. Let us be given $(t_0,y_0),(t'_0,y'_0),(T,k),(T',k')\in\overline{Q}$, where $t_0\leq t'_0$; $|y_0|,|y'_0|\leq\overline{y}_0$; $|k|,|k'|\leq\overline{k}$; $0<\varepsilon\leq T-t_0,T'-t'_0$. Define Π, $\hat{\Pi}$, $\delta\Pi$ as above. Then using the estimates (4.3), (4.12), and the results symmetric in the variables (T,k), and using also well-known results related to $\hat{\Pi}$, which is explicitly given by the Black–Scholes formula, it follows that

$$\begin{aligned}
&|\Pi_{T,k}(t_0,y_0)-\Pi_{T',k'}(t'_0,y'_0)|\\
&\leq|\Pi_{T,k}(t_0,y_0)-\Pi_{T',k'}(t_0,y_0)|+|\Pi_{T',k'}(t_0,y_0)-\Pi_{T',k'}(t'_0,y'_0)|\\
&\leq|\delta\Pi_{T,k}(t_0,y_0)-\delta\Pi_{T',k'}(t_0,y_0)|+|\hat{\Pi}_{T,k}(t_0,y_0)-\hat{\Pi}_{T',k'}(t_0,y_0)|\\
&+|\delta\Pi_{T',k'}(t_0,y_0)-\delta\Pi_{T',k'}(t'_0,y'_0)|+|\hat{\Pi}_{T',k'}(t_0,y_0)-\hat{\Pi}_{T',k'}(t'_0,y'_0)|\\
&\leq\|\delta\Pi_{\cdot}(t_0,y_0)\|_{\mathcal{C}^0_\theta(\overline{Q}_{t_0})}\big(|T-T'|^\theta+|k-k'|^\theta\big)+C^\varepsilon_p(\underline{t},\overline{y}_0,\overline{T},\overline{k};R,\underline{a},\overline{a})(|T-T'|+|k-k'|)\\
&+\|\delta\Pi_{T',k'}(\cdot)\|_{\mathcal{C}^0_\theta(\overline{Q}^{T'})}\big(|t_0-t'_0|^\theta+|y_0-y'_0|^\theta\big)+C^\varepsilon_p(\underline{t},\overline{y}_0,\overline{T},\overline{k};R,\underline{a},\overline{a})(|t_0-t'_0|+|y_0-y'_0|)\\
&\leq C'_p(C_p(\underline{t},\overline{y}_0,\overline{T};R,\underline{a},\overline{a})\vee C_p(\underline{t},\overline{T},\overline{k};R,\underline{a},\overline{a}))\\
&\times\big(|T-T'|^\theta+|k-k'|^\theta+|t_0-t'_0|^\theta+|y_0-y'_0|^\theta\big)\\
&+C^\varepsilon_p(\underline{t},\overline{y}_0,\overline{T},\overline{k};R,\underline{a},\overline{a})\,(|T-T'|+|k-k'|+|t_0-t'_0|+|y_0-y'_0|)\;.
\end{aligned}$$

3. Using (4.12), if ${p'}^{-1} = p^{-1} + \rho^{-1}$, by Hölder's inequality we obtain

$$\begin{aligned}\|h(\partial^2_{y^2} - \partial_y)\Pi\|_{L_{p'}(Q^T)} &\leq \|h\|_{L_\rho(Q^T)}\ \|(\partial^2_{y^2} - \partial_y)\Pi\|_{L_p(Q^T)}\\ &\leq C'_{p'}\ \|h\|_{H^1(Q)}\end{aligned}$$

through the Sobolev embedding in Theorem 3.2(1). Using estimates (4.12) for Π and (4.3) for $d\Pi$ and $d\Pi'$, we obtain similarly

$$\begin{aligned}\|d\Gamma\|_{L_{p''}(Q^T)} &\leq \|h'(\partial^2_{y^2} - \partial_y)d\Pi\|_{L_{p''}(Q^T)}\ +\ \|h(\partial^2_{y^2} - \partial_y)d\Pi'\|_{L_{p''}(Q^T)}\\ &\leq \|h'\|_{L_\nu(Q^T)}\ \|(\partial^2_{y^2} - \partial_y)d\Pi\|_{L_{p'}(Q^T)}\\ &\quad +\ \|h\|_{L_\nu(Q^T)}\ \|(\partial^2_{y^2} - \partial_y)d\Pi'\|_{L_{p'}(Q^T)}\\ &\leq C''_{p''}\ \|h\|_{H^1(Q)}\ \|h'\|_{H^1(Q)}\ .\end{aligned}$$

4. The estimates for $d\Pi$ and $d^2\Pi$ result from point 3 and Theorem 4.2. Let us additionally suppose that $a + h \in \mathcal{M}_Q(\underline{a}, \overline{a})$. By linearity as above, $\varepsilon^{-1}\delta_\varepsilon\Pi$ is the $L_p(Q^T)$-solution of $BS'_{Q^T}(a + \varepsilon h; \Gamma)$, and $\varepsilon^{-1}\delta_\varepsilon\Pi$ converges in $\mathcal{C}^0_\theta(\overline{Q}^T) \cap W^{1,2}_p(Q^T)$, when $\varepsilon \to 0$, towards the solution $d\Pi$ of $BS'_{Q^T}(a; \Gamma)$. Similarly, $\varepsilon^{-1}\delta_\varepsilon d\Pi$ is the $L_p(Q^T)$-solution of $BS'_{Q^T}(a + \varepsilon h; d\Gamma_\varepsilon)$, where

$$\begin{aligned}d\Gamma_\varepsilon \equiv\ & h(\partial^2_{y^2} - \partial_y)d\Pi_{T,k}(\cdot; a).h'\\ &+\ h'(\partial^2_{y^2} - \partial_y)\left[\varepsilon^{-1}(\Pi_{T,k}(\cdot; a + \varepsilon h) - \Pi_{T,k}(\cdot; a))\right]\ .\end{aligned}$$

Moreover, $d\Gamma_\varepsilon$ converges in $L_p(Q^T)$ to $d\Gamma$ when $\varepsilon \to 0$. Therefore, $\varepsilon^{-1}\delta_\varepsilon d\Pi$ converges in $\mathcal{C}^0_\theta(\overline{Q}^T) \cap W^{1,2}_p(Q^T)$ to $d^2\Pi$ when $\varepsilon \to 0$.

5. Having fixed $\varepsilon > 0$, and $2 < p < p' < \overline{p}$, define ρ such that $p^{-1} = {p'}^{-1} + \rho^{-1}$. By (4.12), we can choose a subset $Q_\varepsilon \equiv Q^T \cap \{|y| \leq Y_\varepsilon\}$ such that $\|(\partial^2_{y^2} - \partial_y)\Pi\|_{L_p(Q^c_\varepsilon)} \leq \varepsilon$, where $Q^c_\varepsilon \equiv Q^T \setminus Q_\varepsilon$. By the assumed weak convergence of $a_n - a$ to 0, and by the Sobolev compact embedding (3.2), $a_n - a$ converges to 0 in $L_\rho(Q_\varepsilon)$. Denoting ${\Gamma'}_n \equiv (a_n - a)(\partial^2_{y^2} - \partial_y)\Pi$, it follows, in the same manner as in the proof of Theorem 4.3, that ${\Gamma'}_n$ converges to 0 in $L_p(Q^T)$. The $L_p(Q^T)$-solution $\Pi_n - \Pi$ of $BS'_{Q^T}(a_n; {\Gamma'}_n)$ then converges to 0 in $\mathcal{C}^0_\theta(\overline{Q}^T) \cap W^{1,2}_p(Q^T)$ when $n \to +\infty$ by Theorem 4.2.

Appendix C. Proof of Theorem 5.4. We are going to construct, for $n \in \mathbb{N}^\star$, $a_n \in a_0 + H^1_Q(\underline{a}, \overline{a})$, which takes values in the vicinity of a, such that when $n \to +\infty$, $a_n - a$ converges to 0 weakly in $H^1(Q)$. Hence, by Proposition 5.1(1), $\Pi_\cdot(t_0, y_0; a_n) - \Pi_\cdot(t_0, y_0; a)$ converges to 0 in $\mathcal{C}^0_\theta(\overline{Q}_{t_0}) \cap W^{1,2}_p(Q_{t_0})$. But no subsequence of $a_n - a$ will converge to 0 strongly in $H^1(Q_{t_0})$. Therefore Π or $\Pi_{|\mathcal{F}}$ cannot be continuously invertible around $\widetilde{\Pi} = \Pi(a)$ or $\pi = \Pi_{|\mathcal{F}}(a)$.

Since $\underline{a} < \overline{a}$, and because a is continuous, there exists an open subset $\mathcal{R} \subset Q_{t_0}$ on which $\underline{a} + \varepsilon \leq a$ or $a + \varepsilon \leq \overline{a}$ for some well-chosen $\varepsilon > 0$. Let us assume, for instance, that $a + \varepsilon \leq \overline{a}$ on a rectangle $\mathcal{R} =]t_1, t_2[\times]0, \varepsilon[$, as well as on the union $\mathcal{T}$ of the two equilateral triangles adjacent to the time boundaries of $\mathcal{R}$, with $\mathcal{R} \cup \mathcal{T} \subset Q_{t_0}$. Let us define $a_n - a = u_n$ to be the continuous function on Q such that the following hold:

1. On $\mathcal{R}$, u_n is a continuous function of the space variable y alone, which vanishes at both sides of the space interval $]0, \varepsilon[$ and oscillates between the values 0 and $1/2n$. More precisely, $\partial_y u_n = -1$ or $+1$ on $]0, \varepsilon[$ according to whether $E\{2ny/\varepsilon\}$ is odd or even.

2. On the left and right of $\mathcal{R}$, u_n decreases to 0 at unit speed with respect to the time variable, then vanishes identically.

3. Outside $\mathcal{R} \cup \mathcal{T}$, u_n vanishes identically.

Therefore, u_n vanishes identically outside $\mathcal{R}$, except on a set of measure tending to 0 as $n \to \infty$. Moreover, for every n, we have on Q

$$0 \leq u_n \leq \varepsilon/2n \leq \varepsilon \ , \quad |\partial_t u_n| \leq 1 \ , \quad |\partial_y u_n| \leq 1 \ .$$

So, by construction, $u_n = a_n - a \in H^1(Q)$ and

$$a_n = (a_n - a) + (a - a_0) + a_0 \in a_0 + H^1_Q(\underline{a}, \overline{a}) \ .$$

Moreover, for $n \in \mathbb{N}^\star$, $|\partial_y u_n| \equiv 1$ on $\mathcal{R}$, so that no subsequence of u_n can converge to 0 strongly in $H^1(Q_{t_0})$. But u_n converges to 0 weakly in $H^1(Q)$. Indeed, for any regular test function $\psi(t, y)$, let us define $\phi(y) = \int_{t=t_1}^{t_2} \partial_y \psi \ dt$. Then

$$\int \int_{\mathcal{R}} (\partial_y u_n)\, (\partial_y \psi) \ dydt = \int_{y=0}^{\varepsilon} (\partial_y u_n)\, \phi(y) \ dy = - \int_{y=0}^{\varepsilon} u_n \phi'(y) \ dy$$

by integration by parts. Since $|u_n| \leq \varepsilon/2n$, this converges to 0 when $n \to \infty$. The rest of the verification is straightforward.

Acknowledgments. I am greatly indebted to Henri Berestycki for his enlightening direction of the first stage of this research during a "Stage industriel pour doctorant INRIA" at CAR (Caisse Autonome de Refinancement, Groupe Caisse des Dépôts, Paris; see Crépey [14, Part IV]). Thanks also to Jerôme Busca for kind advice and encouragement throughout the work.

REFERENCES

[1] L. Andersen and R. Brotherton-Ratcliffe, *The equity option volatility smile: An implicit finite difference approach*, J. Comput. Finance, 1 (2) (1997), pp. 5–37.

[2] M. Avellaneda, C. Friedman, R. Holmes, and D. Samperi, *Calibrating volatility surfaces via relative-entropy minimization*, Appl. Math. Finance, 41 (1997), pp. 37–64.

[3] A. Bensoussan and J.-L. Lions, *Applications des Inéquations Variationnelles en Contrôle Stochastique*, Dunod, Paris, 1978.

[4] H. Berestycki, J. Busca, and I. Florent, *Asymptotics and calibration of local volatility models*, Quant. Finance, 2 (2002), pp. 61–69.

[5] A. Binder, H.W. Engl, C.W. Groetsch, A. Neubauer, and O. Scherzer, *Weakly closed nonlinear operators and parameter identification in parabolic equations by Tikhonov regularization*, Appl. Anal., 55 (1994), pp. 13–25.

[6] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Polit. Econ., 81 (1973), pp. 637–659.

[7] J. Bodurtha and M. Jermakyan, *Nonparametric estimation of an implied volatility surface*, J. Comput. Finance, 2 (4) (1999), pp. 29–60.

[8] I. Bouchouev and V. Isakov, *Uniqueness, stability and numerical methods for the inverse problem that arises in financial markets*, Inverse Problems, 15 (1999), pp. R95–R116.

[9] H. Brézis, *Analyse fonctionnelle: Théorie et application*, Coll. Math. Appl. pour la Maîtrise, Masson, Paris, 1983.

[10] L. Caffarelli, M.G. Crandall, M. Kocan, and A. Swiech, *On viscosity solutions of fully nonlinear equations with measurable ingredients*, Comm. Pure Appl. Math., 49 (1996), pp. 365–397.

[11] T. Coleman, Y. Li, and A. Verma, *Reconstructing the unknown volatility function*, J. Comput. Finance, 2 (3) (1999), pp. 77–102.

[12] M. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[13] M.G. Crandall, M. Kocan, and A. Swiech, *L^p-theory for fully nonlinear parabolic equations*, Comm. Partial Differential Equations, 25 (2000), pp. 1997–2053.
[14] S. Crépey, *Contribution à des méthodes numériques appliquées à la Finance et aux Jeux Différentiels*, Ph.D. thesis, Ecole Polytechnique, France, 2001.
[15] S. Crépey, *Tikhonov Regularization and Calibration of a Local Volatility in Finance—Stability, Convergence and Convergence Rates Issues*, CMAP Internal Research Report 474, CMAP, Ecole Polytechnique, Palaiseau, France, 2002.
[16] S. Crépey, *Calibration of the local volatility in a trinomial tree using Tikhonov regularization*, Inverse Problems, 19 (2003), pp. 91–127.
[17] E. Derman and I. Kani, *Riding on a smile*, Risk, 7 (1994), pp. 32–39.
[18] E. Derman, I. Kani, and N. Chriss, *Implied trinomial trees of the volatility smile*, J. Derivatives, 3 (4) (1996), pp. 7–22.
[19] B. Dupire, *Pricing with a smile*, Risk, 7 (1994), pp. 18–20.
[20] N. El Karoui, *Modèles Stochastiques en Finance (Cours de DEA)*, Lecture Notes, Université Paris VI, 1997.
[21] H.W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, The Netherlands, 1996.
[22] H.W. Engl, K. Kunisch, and A. Neubauer, *Convergence rates for Tikhonov regularisation of nonlinear ill-posed problems*, Inverse Problems, 5 (1989), pp. 523–540.
[23] E. Fabes, *Singular integrals and partial differential equations of parabolic type*, Studia Math., 28 (1966), pp. 6–131.
[24] N. Jackson, E. Süli, and S. Howison, *Computation of deterministic volatility surfaces*, J. Comput. Finance, 2 (2) (1999), pp. 5–32.
[25] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, Berlin, 1988.
[26] N.V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, Berlin, 1980.
[27] K. Kunisch and G. Geymayer, *Convergence rates for regularized nonlinear ill-posed problems*, in Modelling and Inverse Problems of Control for Distributed Parameter Systems, Lecture Notes in Control and Inform. Sci. 154, Springer-Verlag, Berlin, 1991, pp. 81–92.
[28] O. Ladyzhenskaya, V. Solonnikov, and N. Ural'tseva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.
[29] R. Lagnado and S. Osher, *A technique for calibrating derivative security pricing models: Numerical solution of an inverse problem*, J. Comput. Finance, 1 (1) (1997), pp. 13–25.
[30] B. Larrouturou and P.L. Lions, *Méthodes mathématiques pour les sciences de l'ingénieur: Optimisation et Analyse Numérique*, Lecture Notes, Ecole Polytechnique, 1994.
[31] R.C. Merton, *Theory of rational option pricing*, Bell J. Econom. and Management Sci., 4 (1973), pp. 141–183.
[32] A. Neubauer, *Tikhonov regularization for nonlinear ill-posed problems: Optimal convergence and finite dimensional approximation*, Inverse Problems, 5 (1989), pp. 541–557.
[33] A. Neubauer and O. Scherzer, *Finite dimensional approximation of Tikhonov regularised solution of nonlinear ill-posed problems*, Numer. Funct. Anal. Optim., 11 (1990), pp. 85–99.
[34] R. Rebonato, *Volatility and Correlation in the Pricing of Equity, FX and Interest-Rate Options*, John Wiley, New York, 1999.
[35] M. Rubinstein, *Implied binomial trees*, J. Finance, 69 (1994), pp. 771–818.
[36] O. Scherzer, *The use of Tikhonov regularization in the identification of electrical conductivities from over determined boundary data*, Results Math., 22 (1992), pp. 598–618.
[37] D.W. Stroock and S.R.S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, 1979.
[38] M. Tikhonov, *Regularization of incorrectly posed problems*, Soviet Math. Dokl., 4 (1963), pp. 1624–1627.
[39] L. Wang, *On the regularity theory of fully nonlinear parabolic equations.* I, Comm. Pure Appl. Math., 45 (1992), pp. 27–76.

SIAM J. MATH. ANAL.
Vol. 34, No. 5, pp. 1207–1225

NEIGHBORHOODS OF PARALLEL WELLS IN TWO DIMENSIONS THAT SEPARATE GRADIENT YOUNG MEASURES*

KEWEI ZHANG†

Abstract. We give estimates for the closed ϵ-neighborhood K_ϵ of the set $K = \cup_{i=1}^k \lambda_i SO(2) \subset M^{2\times 2}$ of multiple parallel elastic wells such that $\text{dist}(Du_j, K_\epsilon) \to 0$ in $L^1(\Omega)$ implies, up to a subsequence, $\text{dist}(Du_j, (\lambda_{i_0} SO(2))_\epsilon) \to 0$ in $L^1(\Omega)$ for some $1 \le i_0 \le k$, where $\Omega \subset \mathbb{R}^2$ is an arcwise connected domain. In other words, K_ϵ separates gradient Young measures.

Key words. two-dimensional parallel wells, separation of gradient Young measures, Laplacian operator, BMO estimate, quasi-convex functions, density argument

AMS subject classifications. 49J45 49N60 73C50 73V25

PII. S0036141001392773

1. Introduction.

The study of weak convergent sequences of gradients approaching a compact set of matrices and their corresponding gradient Young measures [15] is the central theme for the variational approach to material microstructure [8, 9, 21, 5]. An important mathematical question in this approach is the following [5]: Given a compact set $K \subset M^{N\times n}$ of real matrices and a bounded sequence of vector-valued mappings $u_j : \Omega \subset \mathbb{R}^n \to \mathbb{R}^N$ such that the sequence of gradients (Du_j) satisfies $\text{dist}(Du_j, K) \to 0$ in $L^1(\Omega)$ as $j \to \infty$, what can we say about the possible oscillation of (Du_j) (mathematically, the gradient Young measure generated by (Du_j))? In particular, in the multiwell model of material microstructure, by using nonlinear elasticity, we see that the set $K \subset M^{n\times n}$ with $n = 2$ or 3 consists of finitely many copies of $SO(n)$ in the form $K = \cup_{i=1}^k SO(n)H_i$, where H_i's are positive definite matrices. Each $SO(n)H_i$ is called an elastic well.

In practice, one uses the algebraic properties of the set K. If there are rank-one connections in K, one can construct microstructures by using laminates [9, 10, 21] or laminates within laminates [5]. This construction does not give all (mathematically) possible microstructures, as shown by an example due to Šverák [21]. For certain sets K without rank-one connections, one seeks to show that the set prevents the formation of microstructure by using partial differential equation methods [25, 26, 19] or the minors relations [5]. The question we address in this paper, loosely speaking, lies between the two situations above; that is, we give conditions for certain disconnected sets in the multiwell model in two dimensions which prevent "large" scale oscillations among different wells, while microstructures can be formed "locally" near each individual well.

In their study of metastability and local minimizers, Ball and James [9] addressed this problem. They established by a contradiction argument that for a disjoint set $K = K_1 \cup K_2 \subset M^{N\times n}$ with $K_1 \cap K_2 = \emptyset$, which separates gradient Young measures, there is some $\epsilon > 0$ such that the closed ϵ-neighborhood $K_\epsilon = (K_1)_\epsilon \cup (K_1)_\epsilon$ still separates gradient Young measures.

In this paper we give estimates of closed ϵ-neighborhoods K_ϵ of the set of parallel

*Received by the editors July 23, 2001; accepted for publication (in revised form) October 4, 2002; published electronically April 17, 2003.
http://www.siam.org/journals/sima/34-5/39277.html
†School of Mathematical Sciences, University of Sussex, Falmer, Brighton, BN1 9QH, UK (k.zhang@sussex.ac.uk).

multielastic wells $K = \cup_{i=1}^k \lambda_i SO(2) \subset M^{2\times 2}$, $0 < \lambda_1 < \cdots < \lambda_k$ in two dimensions. This is a continuation of the earlier work [32] for finite sets in a subspace of $M^{N\times n}$ without rank-one matrices. The main feature of the present case is that the set K has nontrivial topology, while it is the simplest model among the multielastic well structure $\cup_{i=1}^k SO(2)H_i$ in two dimensions [8, 28, 9, 5, 21]. We have the following theorem.

THEOREM 1. *Let $k \geq 1$, and let $K = \cup_{i=1}^k \lambda_i SO(2) \subset M^{2\times 2}$ be given as above. Suppose $\Omega \subset \mathbb{R}^2$ is a bounded arcwise connected Lipschitz domain. Then there is some $\epsilon_1 > 0$ depending on $r_K = \min_{1\leq i\leq k-1}(\lambda_{i+1}^2 - \lambda_i^2)$, $g_K = \min_{1\leq i\leq k-1}(\lambda_{i+1} - \lambda_i)$, and $d_K = 2\lambda_k$ such that for every $0 < \epsilon \leq \epsilon_1$ and every bounded sequence $(u_j) \subset \mathbb{W}^{1,1}(\Omega, \mathbb{R}^2)$ satisfying*

$$\lim_{j\to\infty} \int_\Omega \operatorname{dist}(Du_j,\, K_\epsilon)dx = 0, \tag{1.1}$$

there is a weak convergent subsequence $u_{j_s} \rightharpoonup u$ in $W^{1,1}(\Omega,\, \mathbb{R}^2)$ and some $1 \leq i_0 \leq k$ such that

$$\lim_{s\to\infty} \int_\Omega \operatorname{dist}(Du_{j_s},\, [\lambda_{i_0} SO(2)]_\epsilon)dx = 0 \quad \textit{and} \quad Du(x) \in [\lambda_{i_0} SO(2)]_\epsilon \ \textit{ a.e. in } \ \Omega. \tag{1.2}$$

Remark 1. Theorem 1 can be stated by using gradient Young measures. Suppose (u_j) satisfies (1.1) and let $\{\nu_x\}_{x\in\Omega}$ be the family of gradient Young measures [15] corresponding to a subsequence of (Du_j). Clearly, the support of ν_x satisfies $\operatorname{supp}\nu_x \subset K_\epsilon$ a.e. Then for some $\epsilon_1 > 0$ depending on the parameters above, we claim that $\operatorname{supp}\nu_x \subset (\lambda_{i_0} SO(2))_\epsilon$ a.e. for some $1 \leq i_0 \leq k$ and the weak limit $Du(x) = \bar\nu_x \in (\lambda_{i_0} SO(2))_\epsilon$ when $0 < \epsilon < \epsilon_1$, where $\bar\nu_x = \int_{K_\epsilon} \lambda d\nu_x$ is the integral average of ν_x.

A disjoint compact set $K = K_1 \cup K_2 \subset M^{N\times n}$ with $K_1 \cap K_2 = \emptyset$ is said to separate gradient Young measures if, for any family of Young measures $\{\nu_x\}_{x\in\Omega}$ supported in K, either $\operatorname{supp}\nu_x \subset K_1$ a.e. or $\operatorname{supp}\nu_x \subset K_2$ a.e. [10]. Our contribution for the present case is a direct estimate of the neighborhood K_ϵ of K that still separates gradient Young measures.

Our approach is based on Schauder's estimates in BMO and Campanato spaces for the Laplacian operator [14], the weak continuity of Jacobians [20, 3, 11], and a recent approximation result due to Müller [22], improving upon an earlier result of the author [30] for sequences of gradients approaching a compact set $K \subset M^{N\times n}$.

Let E_∂ and $E_{\bar\partial}$ be the subspaces of conformal and anticonformal matrices in $M^{2\times 2}$. Note that E_∂ and $E_{\bar\partial}$ are orthogonal complements to each other. We denote by P_{E_∂} and $P_{E_{\bar\partial}}$ the orthogonal projections to these subspaces, respectively. Let $Q\operatorname{dist}^2(A, K)$ be the quasi-convex relaxation of $\operatorname{dist}^2(A, K)$.

Since in our case we can calculate explicitly the quasi-convex relaxation $Q\operatorname{dist}^2(A, K)$, we are able to locate the weak limit Du of Du_j by showing that $Du \in K_\epsilon$ a.e. The use of the homogeneous Young measure [15] makes it possible for us to localize our problem first by considering sequences with fixed affine boundary values. Due to the fact that our set K is contained in E_∂, the projection $P_{E_{\bar\partial}} Du_j$ of the gradient Du_j to its orthogonal complement $E_{\bar\partial}$ is elliptic [5] and the operator $2\operatorname{div} P_{E_{\bar\partial}} Du_j = \Delta u_j$ is exactly the Laplacian. The local Schauder estimate on the approximate solutions v_j obtained in [22] shows that the BMO seminorm of Dv_j is small, so we can use the special geometric and analytic features of the Jacobian, together with a density argument, to establish Theorem 1.

We conclude this section by examining the geometry of the quasi-convex relaxation of the squared distance function to K in Theorem 1. One of the implications of our calculations is that for $0 < \epsilon \le g_K\sqrt{2}/2$, if $\lim_{j\to\infty}\int_\Omega \operatorname{dist}^2(Du_j,\, K_\epsilon)dx = 0$ and u_j converges weakly to u in $W^{1,2}$, then $Du(x) \in K_\epsilon$ a.e. We have the following theorem.

THEOREM 2. *Suppose K is given as in Theorem* 1. *Then the quasi-convex relaxation $Q\operatorname{dist}^2(A, K)$ is given by $Q\operatorname{dist}^2(A, K) = C_{E_\partial}[\operatorname{dist}^2(P_{E_\partial}(A),\, K) + |P_{E_{\bar\partial}}(A)|^2] + [|P_{E_{\bar\partial}}(A)|^2 - |P_{E_\partial}(A)|^2]$, where $C_{E_\partial}[\operatorname{dist}^2(P_{E_\partial}(A),\, K) + |P_{E_\partial}(A)|^2]$ is the convexification of $\operatorname{dist}^2(P_{E_\partial}(A),\, K) + |P_{E_\partial}(A)|^2$ in E_∂. Furthermore,*

(i) *the relaxation is bounded below by the function itself:*

$$Q\operatorname{dist}^2(A,\, K) \ge \frac{1}{2}\operatorname{dist}^2(A,\, K),\ A \in M^{2\times 2};$$

(ii) *whenever* $\operatorname{dist}(A,\, K) \le g_K\sqrt{2}/2$ *with* g_K *given by Theorem* 1, *that is,* $A \in K_{\sqrt{2}g_K/2}$,

$$Q\operatorname{dist}^2(A,\, K) = \operatorname{dist}^2(A,\, K);$$

(iii) *let* $F_\epsilon(X) = \max\{Q\operatorname{dist}^2(A,\, K) - \epsilon^2,\ 0\}$ *for* $0 < \epsilon \le g_K\sqrt{2}/2$; *then* $F_\epsilon \ge 0$ *is a quasi-convex function with quadratic growth and* $F_\epsilon^{-1}(0) = K_\epsilon$.

Theorem 2 shows that at least the quasi-convex relaxation $Q\operatorname{dist}^2(A,\, K)$ does not have any effect on $\operatorname{dist}^2(A,\, K)$ as long as A is in the closed neighborhood $K_{\sqrt{2}g_K/2}$. If a bounded sequence (u_j) in $W^{1,2}(\Omega,\, \mathbb{R}^N)$ converges weakly to u and $\lim_{j\to\infty}\int_\Omega \operatorname{dist}^2(Du_j,\, K_\epsilon)dx = 0$, then $\lim_{j\to\infty}\int_\Omega F_\epsilon(Du_j)dx = 0$, and hence by [2] $\int_\Omega F_\epsilon(Du)dx = 0$, which implies $Du(x) \in K_\epsilon$.

In section 2, notation and preliminaries are given that are needed for proving our main theorem. We establish Theorem 1 in section 3 through two lemmas by assuming Theorem 2. Finally, we prove Theorem 2 in section 4.

2. Preliminaries. Throughout this paper, Ω denotes a bounded arcwise connected open subset of $\mathbb{R}^n$ with Lipschitz boundary. By an arcwise connected domain Ω we mean that for any $x_1, x_2 \in \Omega$, there is a piecewise affine curve $\gamma : [0, 1] \to \Omega$ such that $\gamma(0) = x_1$, $\gamma(1) = x_2$ and each affine piece of γ is parallel to one of the coordinate axes. We denote by $M^{N\times n}$ the space of real $N \times n$ matrices ($N,\, n \ge 2$) with inner product $A \cdot B = \operatorname{tr}(A^TB)$ and norm $|A| = (\operatorname{tr} A^TA)^{1/2}$, where A^T and tr are the transpose of A and the trace operator, respectively. We denote the Lebesgue spaces $L^p(\Omega, \mathbb{R}^N)$ and Sobolev spaces $W^{1,p}(\Omega,\, \mathbb{R}^N)$ and $W_0^{1,p}(\Omega, \mathbb{R}^N)$ for vector-valued functions $u : \Omega \to \mathbb{R}^N$ as usual [1]. The Lebesgue measure of a measurable set S in $\mathbb{R}^n$ is denoted by $\operatorname{meas}(S)$, and we use $\rightharpoonup$ and $\overset{*}{\rightharpoonup}$ to denote weak convergence and weak-$*$ convergence, respectively. The integral average of a (matrix-valued) function f over a measurable set S is written as

$$\fint_S f(x)\,dx = \frac{1}{\operatorname{meas}(S)}\int_S f(x)dx := [f]_S.$$

We define the p-distant function from $Y \in M^{N\times n}$ to a set $K \subset M^{N\times n}$ by $\operatorname{dist}^p(Y, K) := \inf_{A\in K}|Y - A|^p$. The subspaces of conformal and anticonformal matrices are given by

$$E_\partial = \left\{\begin{pmatrix} a & b \\ -b & a\end{pmatrix},\ a,\, b \in \mathbb{R}\right\}, \quad E_{\bar\partial} = \left\{\begin{pmatrix} a & b \\ b & -a\end{pmatrix},\ a,\, b \in \mathbb{R}\right\},$$

respectively. Note that E_∂ and $E_{\bar\partial}$ are orthogonal to each other and

$$SO(2) = \{A \in M^{2\times 2},\ A^T A = I,\ \det A = 1\} \subset E_\partial.$$

A continuous function $f : M^{N\times n} \to \mathbb{R}$ is quasi-convex (see [20, 3]) if

$$\int_U f(A + D\phi(x))\,dx \geq f(A)\,\mathrm{meas}(U)$$

for every $A \in M^{N\times n}$, $\phi \in C_0^\infty(U;\mathbb{R}^N)$, and every open bounded subset $U \subset \mathbb{R}^n$. Also, f is called rank-one convex if, for any A, $B \in M^{N\times n}$ with $\mathrm{rank}(A-B) = 1$ and any $0 \leq \lambda < 1$, $f(\lambda A + (1-\lambda)B) \leq \lambda f(A) + (1-\lambda) f(B)$. It is well known that quasi convexity implies rank-one convexity [20, 3, 11]. However, the converse is not true [27]. It is also well known that the Jacobian $A \to \det(A)$ is quasi-convex.

For a continuous function $f : M^{N\times n} \to \mathbb{R}$ bounded below, the quasi-convex relaxation Qf and rank-one convex relaxation Rf of f are defined, respectively, by $Qf = \sup\{g \leq f,\ g \,\text{quasiconvex}\}$ and $Rf = \sup\{g \leq f,\ g \,\text{rank-one convex}\}$. It is well known that in general $Qf \leq Rf$ (see [11]).

There is an iterative construction of Rf for a given continuous function f due to Kohn and Strang [16, 17, 18], namely,

$$\begin{cases} R_0 f = f, \\ R_{k+1} f(A) = \inf\{\lambda R_k f(A_1) + (1-\lambda) R_k f(A_2), \\ \lambda A_1 + (1-\lambda) A_2 = A, \quad \mathrm{rank}(A_1 - A_2) \leq 1\}. \end{cases} \tag{2.1}$$

It was proved in [16, 17, 18] that $Rf = \lim_{k\to\infty} R_k f$. We call this construction the Kohn–Strang scheme, which will be used to establish Theorem 2. Similarly, we see that the convex envelope Cf can also be calculated by dropping the rank-one restriction in (2.1):

$$\begin{cases} C_0 f = f, \\ C_{k+1} f(A) = \inf\{\lambda C_k f(A_1) + (1-\lambda) R_k f(A_2), \quad \lambda A_1 + (1-\lambda) A_2 = A\}, \\ Cf = \lim_{k\to\infty} C_k f. \end{cases} \tag{2.2}$$

We use the following theorem concerning the existence and properties of Young measures [29, 4, 15] and the homogeneous Young measures [15].

PROPOSITION 1. *Let (z_j) be a bounded sequence in $L^1(\Omega;\mathbb{R}^s)$. Then there exist a subsequence (z_{j_k}) of $(z^{(j)})$ and a family $(\nu_x)_{x\in\Omega}$ of probability measures on $\mathbb{R}^s$, depending measurably on $x \in \Omega$, such that*

$$f(z_{j_k}) \rightharpoonup \int_{\mathbb{R}^s} f(\lambda) d\nu_x(\lambda) \qquad \text{in } L^1(\Omega) \text{ as } k \to \infty$$

for every continuous function $f : \mathbb{R}^s \to \mathbb{R}$ such that $(f(z_{j_k}))$ is sequentially weakly relatively compact in $L^1(\Omega)$.

If the sequence z_j is in the form $z_j = Du_j$, where $\Omega \subset \mathbb{R}^n$ is open and bounded, and (u_j) is a bounded sequence in $W^{1,p}(\Omega, \mathbb{R}^N)$ for some $1 < p \leq \infty$, then the corresponding family of Young measures (ν_x) is called p-gradient Young measures (see [15, 5]). A family of (gradient) Young measures is *trivial* if ν_x is a Dirac measure for almost every x. In this case there exists a function u such that ν_x is the Dirac measure at $Du(x)$, and, up to a subsequence, $Du_k \to Du$ a.e.

The following result on homogeneous Young measures was obtained in [15].

PROPOSITION 2. *Let $\{\nu_x\}_{x\in\Omega}$ be a family of p-gradient Young measures with*

$$\int_{M^{N\times n}} \lambda d\nu_x(\lambda) = Du(x)$$

and $\operatorname{supp}\nu_x \subset K$ *for almost every* $x \in \Omega$ *for a compact set* $K \subset M^{N\times n}$. *Then for almost every* $x_0 \in \Omega$, *there exists a bounded sequence* (ϕ_k) *in* $W_0^{1,\infty}(D,\mathbb{R}^N)$ *such that the corresponding gradient Young measures* $\{\hat\nu_y\}$ *of the sequence* $(Du(x_0)+D\phi_k)$ *satisfy* $\hat\nu_y = \nu_{x_0}$ *for almost every* $y \in D$, *where* D *is the unit open cube in* $\mathbb{R}^n$. *We call* $\nu := \hat\nu_y$ *a homogeneous Young measure.*

Now we recall some definitions and results for linear elliptic systems with constant coefficients [14]. The Campanato spaces $\mathcal{L}^{p,\lambda}(\Omega)$ for $p \ge 1$, $\lambda \ge 1$ on a Lipschitz domain are defined by

$$\mathcal{L}^{p,\lambda}(\Omega) = \left\{ u \in L^p(\Omega),\ \sup_{x_0\in\Omega,\,0<\rho\le\operatorname{diam}(\Omega)} \rho^{-\lambda}\int_{\Omega(x_0,\rho)} |u-[u]_{x_0,\rho}|^p dx = [u]^p_{\mathcal{L}^{p,\lambda}(\Omega)} < \infty \right\},$$

where $\Omega(x_0,\rho) = \Omega\cap B_\rho(x_0)$ and $[u]_{x_0,\rho} = \fint_{\Omega(x_0,\rho)} u dx$.

We also have the local version of the space $BMO(\Omega)$ as L^1 functions on Ω with seminorm

$$\|u\|_{BMO(\Omega)} = \sup\left\{ \left(\fint_Q |u-[u]_Q|^p dx\right)^{1/p},\quad Q\subset\Omega \right\} < +\infty,$$

where $1 \le p < \infty$, $Q \subset \Omega$ are closed cubes with edges parallel to the coordinate axes, and $[u]_Q = \fint_Q u dx$. In this paper we mainly consider BMO on a cube or a ball.

It is well known [14] from John–Nirenberg's inequality that for all $1 \le p < \infty$, the BMO seminorms are equivalent, and one can replace cubes $Q \subset \Omega$ by balls and the resulting seminorm is still equivalent. It is also known that $\mathcal{L}^{p,n}(\Omega)$ is equivalent to $BMO(\Omega)$.

PROPOSITION 3 (see [14, Chap. 3–4]). *Let* $\Omega \subset \mathbb{R}^n$ *be open. Suppose* $u \in W^{1,2}_{loc}(\Omega)$ *is a weak solution of the Poisson equation* $\Delta u = \operatorname{div} f$ *in* Ω *with* $f \in L^\infty(\Omega,\mathbb{R}^n)$; *then for any* $x_0 \in \Omega$ *and* $0 < \rho < R$ *such that* $B_\rho(x_0) \subset B_R(x_0) \subset \bar B_R(x_0) \subset \Omega$, *we have that*

$$\fint_{B_\rho(x_0)} |Du-[Du]_{x_0,\rho}|^2 dx \le C\left[\left(\frac{\rho}{R}\right)^\tau \fint_{B_R(x_0)} |Du-[Du]_{x_0,R}|^2 dx + [f]^2_{\mathcal{L}^{2,2}(\Omega)}\right],$$

where $C > 0$ *and* $0 < \tau < 2$ *are constants.*

Next we state the approximation result of Müller [22], which we will need later.

PROPOSITION 4. *Let* $\Omega \subset \mathbb{R}^n$ *be an open set and let* $K \subset M^{N\times n}$ *be compact and convex. Suppose* $(u_j) \subset W^{1,p}(\Omega,\mathbb{R}^N)$, $1 \le p < \infty$, *and* $\lim_{j\to\infty}\int_\Omega \operatorname{dist}^p(Du_j,K)dx \to 0$. *Then there exists a sequence* (v_j) *of Lipschitz mappings such that*

$$\|\operatorname{dist}(Dv_j,K)\|_{L^\infty} \to 0,\quad \operatorname{meas}\{x\in\Omega,\ u_j \ne v_j\} \to 0 \qquad \text{as } j\to\infty.$$

We conclude this section by briefly recalling the notion of density for a measurable subset V of $\mathbb{R}^n$. A point $x \in \mathbb{R}^n$ is a point of density 1 of V if $\lim_{r\to 0}\operatorname{meas}(B_r(x)\cap$

$V)/\operatorname{meas}(B_r(x) = 1$ and a point of density 0 of V if $\lim_{r\to 0}\operatorname{meas}(B_r(x)\cap V)/\operatorname{meas}(B_r(x)) = 0$ [24, 12]. It is well known [24, 12] that

$$\lim_{r\to 0}\frac{\operatorname{meas}(B_r(x)\cap V)}{\operatorname{meas}(B_r(x))} = 1 \text{ for almost every } x\in V$$

and

$$\lim_{r\to 0}\frac{\operatorname{meas}(B_r(x)\cap V)}{\operatorname{meas}(B_r(x))} = 0 \text{ for almost every } x\in \mathbb{R}^n\setminus V;$$

that is, almost every $x\in V$ is a point of density 1 of V and almost every $x\notin V$ is a point of density 0 of V. It is also known that the balls $B_r(x)$ in the definition can be replaced by cubes $Q_r(x)$ centered at x with edges parallel to one of the coordinate axes and with radius $r>0$ [24].

3. Proof of Theorem 1. We decompose the proof of Theorem 1 into two lemmas. By using homogeneous gradient Young measures [15] in Lemma 1, we localize our problem to a simpler one. We show that if a sequence of gradients Dv_j corresponds to a homogeneous Young measure ν satisfying $\operatorname{supp}\nu\subset K_\epsilon$, then $\operatorname{supp}\nu\subset(\lambda_{i_0}SO(2))_\epsilon$ for some $1\le i_0\le k$. Then in Lemma 2 we deal with the regularity problem that $Du\in K_\epsilon$ a.e. implies that $Du\in(\lambda_{i_0}SO(2))_\epsilon$ for some i_0 a.e. Let $D\subset\mathbb{R}^2$ be the unit closed square $[0,\,1]^2$.

LEMMA 1. *Let $K=\cup_{i=1}^k\lambda_i SO(2)$ with $0<\lambda_1<\cdots<\lambda_k$. Then there is some $\epsilon_2>0$ depending on r_K, g_K, and d_K in Theorem 1 such that for $0<\epsilon\le\epsilon_2$, $A\in(\lambda_{i_0}SO(2))_\epsilon$ with a fixed $1\le i_0\le k$, and $\phi_j\in W_0^{1,\infty}(D,\mathbb{R}^2)$ satisfying $\phi_j\overset{*}{\rightharpoonup}0$ in $W^{1,\infty}(D,\mathbb{R}^2)$ such that $(A+D\phi_j)$ generates the homogeneous gradient Young measure ν with $\operatorname{supp}\nu\subset K_\epsilon$; then $\operatorname{supp}\nu\subset(\lambda_{i_0}SO(2))_\epsilon$.*

The assumption that A is in one of the wells is guaranteed by Theorem 2.

LEMMA 2. *Let K be as in Lemma 1 and let $\Omega\subset\mathbb{R}^n$ be a bounded arcwise connected Lipschitz domain. Then there are some $\epsilon_3>0$ depending on r_K, g_K, and d_K as above such that $u\in W^{1,\infty}(\Omega,\mathbb{R}^N)$, $Du(x)\in K_\epsilon$ for almost every $x\in\Omega$, and $0<\epsilon\le\epsilon_3$ imply $Du(x)\in(\lambda_{i_0}SO(2))_\epsilon$ a.e. in Ω for some $1\le i_0\le k$.*

We prove Lemma 1 first, followed by the proof of Lemma 2. Then the proof of Theorem 1 will follow easily from them.

Before we establish Lemma 1, let me explain the main idea and steps of the proof.

By using induction, Theorem 2, and Proposition 4, we may find another sequence v_j bounded in $W^{1,\infty}$ such that $A+Dv_j$ is in a small neighborhood of $C(K)$, while A is near $\lambda_{i_0}SO(2)$ for some $1\le i_0\le k$. By using BMO seminorm locally on a fixed small square Q_0, we can show that $\|Dv_j\|_{BMO(Q_0)}$ is small.

To deal with the geometry of the set K, we consider $\|\det(A+Dv_j)\|_{BMO(Q_0)}$, which is also small, while the values of $\det(A+Dv_j)$ will be close to the ordered set $\{\lambda_i^2\}$, that is,

$$\det(A+Dv_j)=\sum_{i=1}^k\lambda_i^2\chi_{U_j^i}+\det(A+Dv_j)\chi_{W_j}+O(\epsilon),$$

with U_j^i the subset in Q_0, and where $\det(A+Dv_j)$ is close to λ_i^2, while W_j is the transition part whose measure tends to zero as $j\to\infty$.

Then we consider two cases, either (a) $\lambda_k^2 - \lambda_{i_0}^2 \geq \lambda_{i_0}^2 - \lambda_1^2$, or (b) $\lambda_{i_0}^2 - \lambda_1^2 \geq \lambda_k^2 - \lambda_1^2$. For case (a), we show that $\operatorname{meas}(U_j^k) \to 0$ to finish the proof by the induction assumption. If (b) happens, we can prove that $\operatorname{meas}(U_j^1) \to 0$, and again the proof will be finished.

Under assumption (a), if we let $\alpha_j^k = \operatorname{meas}(U_j^k \cap Q)/\operatorname{meas}(Q)$ for $Q \subset Q_0$, we use the smallness of the *BMO* seminorm of the Jacobian to show that (see (3.8))

$$\alpha_j^k(1-\alpha_j^k) \leq C\left(\epsilon + \fint_Q (\lambda_k^2 + |\det(A+Dv_j)|)\chi_{W_j}dx\right).$$

On the other hand, on Q_0, we will see that $\alpha_j^k < 3/4$ for large j, while at each point $x \in Q_0$ of density 1 for U_j^k, we can find a small square $Q \subset Q_0$ containing x such that $\alpha_j^k > 3/4$. By a continuous deformation of squares, we can find a square Q_x lying between Q and Q_0 over which $\alpha_j^k = 3/4$. The idea here is to "maximize" the left-hand side of the above inequality.

If we substitute this square in the above inequality (i.e., (3.8) below) and assume $\epsilon > 0$ small, we can bound $\operatorname{meas}(Q_x)$ by $\int_{Q_x \cap W_j}(\lambda_k^2 + |\det(A+Dv_j)|)dx$. We then apply Besicovitch's lemma to show that $\operatorname{meas}(U_j^k)$ is bounded by $\int_{W_j}(\lambda_k^2 + |\det(A+Dv_j)|)dx$, which goes to zero; hence $\operatorname{meas}(U_j^k) \to 0$.

Proof of Lemma 1. We use induction. When $k = 1$, there is nothing to prove except that the weak limit satisfies $Du(x) \in (\lambda_{i_0} SO(2))_\epsilon$, which can be checked by using the estimates in Theorem 2 and the weak lower semicontinuity theorem of Acerbi and Fusco [2].

Suppose Lemma 1 is true for $k-1 \geq 1$ and that we seek to prove that it is still true for k. Let $A \in (\lambda_{i_0} SO(2))_\epsilon$. We apply Proposition 4 to $u_j = \phi_j$ and the compact set $C(K_\epsilon) - A$ to obtain a sequence $v_j \in W^{1,\infty}$ such that $Dv_j(x) \in (C(K_\epsilon) - A)_\epsilon$ and $\operatorname{meas}(\{x \in D,\, \phi_j \neq v_j\}) \to 0$. Thus $\int_D |D\phi_j - Dv_j|dx \to 0$ as $j \to \infty$. Letting $h_j = P_{E_{\bar\partial}} Dv_j$, we have $\|h_j\|_{L^\infty} \leq 3\epsilon$ and v_j satisfies $\operatorname{div} P_{E_{\bar\partial}} Dv_j = \operatorname{div} h_j$ in the weak sense. However, we see that $2\operatorname{div} P_{E_{\bar\partial}} Dv_j = \Delta v_j$ (see [6, 7]). Thus in the weak sense, $\Delta v_j = 2\operatorname{div} h_j$ in Ω.

From Schauder estimates for the Laplacian operator (Proposition 3), for a fixed $x_0 \in D$, $0 < \rho < R$ such that $B_\rho(x_0) \subset B_R(x_0) \subset \bar B_R(x_0) \subset B_{2R}(x_0) \subset D$, we have

$$\fint_{B_\rho(x_0)} |Dv_j - [Dv_j]_{x_0,\rho}|^2 dx \leq C\left[\left(\frac{\rho}{R}\right)^\tau \fint_{B_R(x_0)} |Dv_j - [Dv_j]_{x_0,R}|^2 dx + [h_j]^2_{\mathcal{L}^{2,2}(D)}\right],$$

where $C > 0$ and $0 < \tau < 2$ are constants. Hence we have

$$\begin{aligned}\|Dv_j\|^2_{BMO(Q_0)} &\leq C\left[\left(\frac{\rho}{2R}\right)^\tau \fint_{B_{2R}(x_0)} |Dv_j - [Dv_j]_{x_0,2R}|^2 dx + [h_j]^2_{\mathcal{L}^{2,2}(D)}\right] \\ &\leq C\left[\left(\frac{\rho}{2R}\right)^\tau 4(\lambda_k + 3\epsilon)^2 + \epsilon^2\right],\end{aligned}$$

where Q_0 is a cube centered at x_0 with side length ρ. Here we have used the fact that

$\|Dv_j\|_{L^\infty} \le 2\lambda_k+6\epsilon$. Now we choose $\rho>0$ small enough such that $(\frac{\rho}{2R})^\tau 4(\lambda_k+3\epsilon)^2 \le \epsilon^2$. Thus we have, for small $\rho>0$, $\|Dv_j\|^2_{BMO(Q_0)} \le C\epsilon^2$ for all $j>0$. Now we write $A+Dv_j(x)$ as

$$A+Dv_j(x)=\sum_{i=1}^{k} R_j(x)(\lambda_i I+f_j^i(x))\chi_{U_j^i}+(A+Dv_j(x))\chi_{W_j}, \tag{3.1}$$

where $U_j^i=\{x\in Q_0,\, A+Dv_j(x)\in(\lambda_i SO(2))_{2\epsilon}\}$, $i=1,\dots,k$, $W_j=Q_0\setminus(\cup_{i=1}^k U_j^i)$, and $\chi_{U_j^i}$ and χ_{W_j} are the characteristic functions of these sets, respectively. Also $R_j: Q_0\to SO(2)$ is a measurable mapping and f_j^i a small matrix-valued mapping for $i=1,2,\dots,k$, $j=1,2,\dots$. Note that $\mathrm{meas}(W_j)\to 0$ as $j\to\infty$. Let $w_j=\sum_{i=1}^k f_j^i\chi_{U_j^i}$; then w_j is a matrix-valued function with $\|w_j\|_{L^\infty}\le 2\epsilon$. Since A is a constant matrix, we have $\|A+Dv_j\|_{BMO(Q_0,M^{2\times 2})}\le C\epsilon$, and for each cube $Q\subset Q_0$,

$$⨍_Q |A+Dv_j-[A+Dv_j]_Q|dx\le C_1\epsilon, \tag{3.2}$$

where $C_1>0$ is an absolute constant independent of λ_k. We then have, from (3.2) and Taylor's expansion of the Jacobian, that

$$\begin{aligned}
&⨍_Q |\det(A+Dv_j)-[\det(A+Dv_j)]_Q|dx\\
&=⨍_Q\left|\int_0^1 \{\mathrm{adj}(t(A+Dv_j))+(1-t)[(A+Dv_j)]_Q\}\,dt((A+Dv_j)-[(A+Dv_j)]_Q)\right.\\
&\left.-\left[\int_0^1 \{\mathrm{adj}(t(A+Dv_j))+(1-t)[(A+Dv_j)]_Q\}\,dt((A+Dv_j)-[(A+Dv_j)]_Q)\right]_Q\right|dx\\
&\le 4(\lambda_k+3\epsilon)⨍_Q |A+Dv_j-[A+Dv_j]_Q|dx\le 32\lambda_k C_1\epsilon\le C_2\epsilon,
\end{aligned} \tag{3.3}$$

where $\mathrm{adj}(A)$ is the adjoint of A. Note that $C_2>0$ depends on the diameter $2\lambda_k$ of K. Also (3.1) implies

$$\begin{aligned}
\det(A+Dv_j(x))&=\sum_{i=1}^k(\det(R_j(x)\lambda_i I+f_j^i(x)))\chi_{U_j^i}+\det(A+Dv_j(x))\chi_{W_j}\\
&=\sum_{i=1}^k\lambda_i^2\chi_{U_j^i}+\det(A+Dv_j(x))\chi_{W_j}+H_j(x),
\end{aligned}$$

where $|H_j(x)|\le C\epsilon$. Now we substitute this decomposition into (3.3). For each fixed $Q\subset Q_0$,

$$\begin{aligned}
⨍_Q\Bigg|&\sum_{i=1}^k\lambda_i^2\chi_{U_j^i}+\det(A+Dv_j(x))\chi_{W_j}+H_j(x)\\
&-\left[\sum_{i=1}^k\lambda_i^2\chi_{U_j^i}+\det(A+Dv_j)\chi_{W_j}+H_j\right]_Q\Bigg|dx\le C_2\epsilon.
\end{aligned} \tag{3.4}$$

The left-hand side of (3.4) gives

$$
\begin{aligned}
(3.5)\quad & \fint_Q \left| \sum_{i=1}^{k} (\lambda_i^2 \chi_{U_j^i}) + \det(A + Dv_j(x)) \chi_{W_j} + H_j(x) \right. \\
& \qquad \left. - \left[\sum_{i=1}^{k} (\lambda_i^2 \chi_{U_j^i}) + \det(A + Dv_j) \chi_{W_j} + H_j \right]_Q \right| dx \\
& \geq \fint_Q \left| \sum_{i=1}^{k} ((\lambda_i^2 - \lambda_{i_0}^2) \chi_{U_j^i}) - \left[\sum_{i=1}^{k} ((\lambda_i^2 - \lambda_{i_0}^2) \chi_{U_j^i}) \right]_Q \right| dx - 2C\epsilon \\
& \quad - \fint_Q \left| \det(A + Dv_j(x)) \chi_{W_j} - \lambda_{i_0}^2 \chi_{U_j^{i_0}} - \left[\det(A + Dv_j(x)) \chi_{W_j} - \lambda_{i_0}^2 \chi_{U_j^{i_0}} \right]_Q \right| dx.
\end{aligned}
$$

Let $\alpha_j^i = \operatorname{meas}(U_j^i \cap Q)/\operatorname{meas}(Q)$ and $\beta_j = \operatorname{meas}(W_j \cap Q)/\operatorname{meas}(Q)$ so that $(\sum_{i=1}^{k} \alpha_j^i) + \beta_j = 1$ and $\alpha_j^i \geq 0$, $\beta_j \geq 0$, $i = 1, \ldots, k$.

We may assume that either (a) $\lambda_k^2 - \lambda_{i_0}^2 \geq \lambda_{i_0}^2 - \lambda_1^2$, or (b) $\lambda_{i_0}^2 - \lambda_1^2 \geq \lambda_k^2 - \lambda_1^2$.

If (a) holds, from (3.5) we have first that

$$
\begin{aligned}
& \fint_Q \left| \sum_{i=1}^{k} ((\lambda_i^2 - \lambda_{i_0}^2) \chi_{U_j^i}) - \left[\sum_{i=1}^{k} ((\lambda_i^2 - \lambda_{i_0}^2) \chi_{U_j^i}) \right]_Q \right| dx \\
& = \fint_Q \sum_{i=1}^{k} \left| (\lambda_i^2 - \lambda_{i_0}^2) - \left[\sum_{i=1}^{k} ((\lambda_i^2 - \lambda_{i_0}^2) \chi_{U_j^i}) \right]_Q \right| \chi_{U_j^i} dx \\
& \geq \fint_Q \left| (\lambda_k^2 - \lambda_{i_0}^2) - \left[\sum_{i=1}^{k} ((\lambda_i^2 - \lambda_{i_0}^2) \chi_{U_j^i}) \right]_Q \right| \chi_{U_j^k} dx \\
& \geq \fint_Q \left| \lambda_k^2 - \sum_{i=1}^{k} \lambda_i^2 \alpha_j^i \right| \chi_{U_j^k} dx - \beta_j \lambda_{i_0}^2 \\
& \geq \fint_Q \left| \lambda_k^2 - \sum_{i=1}^{k} \lambda_i^2 \alpha_j^i \right| \chi_{U_j^k} dx - \epsilon
\end{aligned}
$$

for large j as $\beta_j \to 0$. Thus,

$$
\begin{aligned}
& \frac{1}{\operatorname{meas}(Q)} \int_{U_j^k} \left| \lambda_k^2 - \sum_{i=1}^{k} \lambda_i^2 \alpha_j^i \right| dx = \alpha_j^k \left(\left| \lambda_k^2 - \sum_{i=1}^{k} \lambda_i^2 \alpha_j^i \right| \right) \\
(3.6)\qquad & \leq C_3 \epsilon + 2 \fint_Q (\lambda_k^2 + |\det(A + Dv_j(x))|) \chi_{W_j} dx \\
& \leq C_3 \epsilon + 2 \fint_Q (\lambda_k^2 + |\det(A + Dv_j(x))|) \chi_{W_j} dx.
\end{aligned}
$$

On the other hand, we have from (3.6) that

(3.7)
$$\alpha_j^k\left(\left|\lambda_k^2-\sum_{i=1}^k\lambda_i^2\alpha_j^i\right|\right)=\alpha_j^k\left(\left(\sum_{i=1}^k(\lambda_k^2-\lambda_i^2)\alpha_j^i\right)+\beta_j\lambda_k^2\right)$$
$$\geq\alpha_j^k\left(\sum_{i=1}^{k-1}(\lambda_k^2-\lambda_{k-1}^2)\alpha_j^i\right)+\alpha_j^k\beta_j\lambda_k^2=\alpha_j^k\left[(\lambda_k^2-\lambda_{k-1}^2)(1-\alpha_j^k-\beta_j)+\beta_j\lambda_k^2\right]$$
$$=(\lambda_k^2-\lambda_{k-1}^2)\alpha_j^k(1-\alpha_j^k)+\alpha_j^k\beta_j\lambda_{k-1}^2\geq(\lambda_k^2-\lambda_{k-1}^2)\alpha_j^k(1-\alpha_j^k)\geq r_K\alpha_j^k(1-\alpha_j^k).$$

Combining (3.6) and (3.7) we obtain

(3.8)
$$\frac{\operatorname{meas}(U_j^k\cap Q)}{\operatorname{meas}(Q)}\left(1-\frac{\operatorname{meas}(U_j^k\cap Q)}{\operatorname{meas}(Q)}\right)\leq C_4\epsilon+\frac{2}{r_K}⨍_Q(\lambda_k^2+|\det(A+Dv_j(x))|)\chi_{W_j}dx.$$

Here $C_4>0$ depends on λ_k^2 and r_K.

We also have $A=R_0(\lambda_{i_0}I+A_0)$ with A_0 a symmetric matrix, $|A_0|\leq\epsilon$, and $R_0\in SO(2)$; hence

$$|\det(A)-\lambda_{i_0}^2|\leq C(\lambda_k\epsilon+\epsilon^2),$$

where $C>0$ is an absolute constant. Now since $\det(A+Dv_j)\rightharpoonup\det(A)$ in $L^p(\Omega)$, $1\leq p<\infty$, we have for large j that

$$\left|⨍_{Q_0}(\det(A+Dv_j)-\det(A))dx\right|\leq\epsilon,$$

so that

$$\left|⨍_{Q_0}(\det(A+Dv_j)-\det(A)+(\lambda_{i_0}^2-\lambda_1^2)dx\right|\leq\epsilon+(\lambda_{i_0}^2-\lambda_1^2). \tag{3.9}$$

Therefore

$$\begin{aligned}&\left|⨍_{Q_0}\left(\sum_{i=1}^k(\lambda_i^2-\lambda_{i_0}^2)\chi_{U_j^i}+(\lambda_{i_0}^2-\lambda_1^2)\right)dx\right|\\&\leq C_2(\epsilon+\epsilon^2)+(\lambda_{i_0}^2-\lambda_1^2)+2⨍_{Q_0}(\lambda_k^2+|\det(A+Dv_j)|)\chi_{W_j}dx,\end{aligned} \tag{3.10}$$

while

$$\begin{aligned}&\left|⨍_{Q_0}\left(\sum_{i=1}^k(\lambda_i^2-\lambda_{i_0}^2)\chi_{U_j^i}+(\lambda_{i_0}^2-\lambda_1^2)\right)dx\right|\\&=\left|⨍_{Q_0}\left(\sum_{i=1}^k(\lambda_i^2-\lambda_1^2)\chi_{U_j^i}\right)+(\lambda_{i_0}^2-\lambda_1^2)\chi_{W_j}dx\right|\\&\geq(\lambda_k^2-\lambda_1^2)⨍_{Q_0}\chi_{U_j^k}dx=(\lambda_k^2-\lambda_1^2)\frac{\operatorname{meas}(U_j^k\cap Q_0)}{\operatorname{meas}(Q_0)}.\end{aligned} \tag{3.11}$$

Consequently,

$$\frac{\operatorname{meas}(U_j^k \cap Q_0)}{\operatorname{meas}(Q_0)} \le C_5\epsilon + \frac{\lambda_{i_0}^2 - \lambda_1^2}{\lambda_k^2 - \lambda_1^2} + \frac{2}{\lambda_k^2 - \lambda_1^2} \fint_{Q_0} (\lambda_k^2 + |\det(A + Dv_j)|)\chi_{W_j} dx$$
$$\le C_5\epsilon + \frac{1}{2} + \frac{2}{\lambda_k^2 - \lambda_1^2} \fint_{Q_0} (\lambda_k^2 + |\det(A + Dv_j)|)\chi_{W_j} dx,$$

where $C_5 > 0$ is a constant depending on r_K and d_K. Here we have used assumption (a), which gives $(\lambda_{i_0}^2 - \lambda_1^2)/(\lambda_k^2 - \lambda_1^2) \le 1/2$. Since $\fint_{Q_0} (\lambda_k^2 + |\det(A+Dv_j)|)\chi_{W_j} dx \to 0$ as $j \to \infty$, we have

$$\frac{\operatorname{meas}(U_j^k \cap Q_0)}{\operatorname{meas}(Q_0)} < \frac{3}{4} \qquad \text{if we require } \; C_5\epsilon < \frac{1}{4} \tag{3.12}$$

for large $j > 0$.

If (b) holds, we may obtain estimates similar to those above by replacing $\lambda_{i_0}^2 - \lambda_1^2$ by $\lambda_{i_0}^2 - \lambda_k^2$ in (3.10).

Now for any $x \in U_j^k$ with density 1, it is easy to see that x is an interior point of Q_0. We define a function on squares Q_x containing x with $Q_x \subset Q_0$ (by a square we always mean an open square in $\mathbb{R}^2$ with its edges parallel to the coordinate axes):

$$G(Q_x) = \frac{\operatorname{meas}(U_j^k \cap Q_x)}{\operatorname{meas}(Q_x)}.$$

Then the function $G(Q_x)$ is continuous with respect to continuous deformations of $Q_x \subset Q_0$ when $\operatorname{meas}(Q_x) > 0$. If we take $Q_x = Q_0$, then $G(Q_0) < 3/4$. We can also find some $Q_x^* \subset Q_0$ strictly inside Q_0 such that $G(Q_x^*) > 3/4$. Therefore, starting from Q_0 we may find a family of decreasing squares $Q(t) \subset Q_0$, $0 \le t \le 1$, such that $Q(t) \subset Q(s)$ if $t > s$, $\operatorname{meas}(Q(s) \setminus Q(t)) \to 0$ as $t \to s$ or $s \to t$, and $Q(0) = Q_0$, $Q(1) = Q_x^*$. Since $G(Q(t))$ is a continuous function of t, the intermediate value theorem implies that we may find some $Q(t_0) := Q_x$ containing x and inside Q_0 such that $G(Q_x) = 3/4$. Substituting this Q_x in (3.8), we obtain

$$G(Q_x)(1 - G(Q_x)) \le C_4\epsilon + \frac{2}{r_K} \fint_{Q_x} \left(\lambda_k^2 + |\det(A + Dv_j(y))|\right) \chi_{W_j} dy.$$

Substituting $G(Q_x) = 3/4$ in the last inequality, we obtain

$$\frac{3}{16} \le C_4\epsilon + \frac{2}{r_K} \fint_{Q_x} \left(\lambda_k^2 + |\det(A + Dv_j(y))|\right) \chi_{W_j} dy,$$

so that

$$\left(\frac{3}{16} - C_4\epsilon\right) \operatorname{meas}(Q_x) \le \frac{2}{r_K} \int_{Q_x} \left(\lambda_k^2 + |\det(A + Dv_j(y))|\right) \chi_{W_j} dy. \tag{3.13}$$

Here we need the left-hand side of (3.13) to be positive, so we require $\epsilon < 3/(16C_4)$. For each $x \in U_j^k$ with density 1, there is some $Q_x \subset Q_0$ containing x with the side-length less than 1 because $Q_0 \subset D$. Hence $\{Q_x\}$ covers U_j^k except for a subset of measure zero. From Besicovitch's covering lemma (see, for example, [12, 24, 14]), there is a countable subcollection $\{Q_s\}_{s=1}^{\infty}$ of $\{Q_x\}$ which still covers the subset of U_j^k

of points with density 1, with Q_s's overlapping at most 4 times (in two-dimensional Euclidean spaces), or $\sum_s^\infty \chi_{Q_s} \le 4$. Now we replace Q_x by Q_s and sum up the inequality over all Q_s's to obtain

$$\begin{aligned}\left(\frac{3}{16} - C_4\epsilon\right)\operatorname{meas}(U_j^k) &\le \sum_{s=1}^{\infty} \frac{2}{r_K}\int_{Q_s\cap W_j} \left(\lambda_k^2 + |\det(A+Dv_j(y))|\right) dy \\ &\le 8\frac{1}{r_K}\int_{W_j}\left(\lambda_k^2 + |\det(A+Dv_j(y))|\right) dy.\end{aligned} \tag{3.14}$$

So there is a $\gamma > 0$ such that for large $j > 0$,

$$\operatorname{meas}(U_j^k) \le \gamma \int_{W_j} \left(\lambda_k^2 + |\det(A+Dv_j(y))|\right)\chi_{W_j} dy.$$

We see that $\operatorname{meas}(U_j^k) \to 0$ as $j \to \infty$ if (a) holds, because

$$\lim_{j\to\infty}\int_{W_j}\left(\lambda_k^2 + |\det(A+Dv_j(y))|\right) dy = 0.$$

Hence $\operatorname{dist}^2(A+Dv_j, \cup_{i=1}^{k-1}(\lambda_k SO(2))_\epsilon) \to 0$ in $L^p(Q_0)$ as $j\to\infty$ for each fixed $1 \le p < \infty$. Therefore the gradient Young measure $\{\nu_x\}_{x\in D}$ corresponding to $A+Dv_j$ satisfies $\operatorname{supp}\nu_x \subset \cup_{i=1}^{k-1}(\lambda_i SO(2))_\epsilon$ a.e., at least for $x \in Q_0$. However, since $|Dv_j - D\phi_j| \to 0$ in $L^2(D)$, $A+Dv_j$ and $A+D\phi_j$ correspond to the same Young measures, while the Young measure generated by $A+D\phi_j$ is ν—a homogeneous Young measure. Thus $\nu = \nu_x$ for almost every $x \in D$, which implies that $\operatorname{supp}\nu \subset \cup_{i=1}^{k-1}(\lambda_i SO(2))_\epsilon$. From the induction assumption, $\operatorname{supp}\nu \subset (\lambda_{i_0} SO(2))_\epsilon$.

If (b) happens, we can show that $\operatorname{meas}(U_j^1) \to 0$ so that $\operatorname{supp}\nu \subset \cup_{i=2}^{k}(\lambda_i SO(2))_\epsilon$. In either case, the proof of Lemma 1 is finished. □

Next we prove Lemma 2. The idea is similar to that of Lemma 1, with a few exceptions. We write $Du(x) = \sum_{i=1}^k R(x)(\lambda_1 I + f^i(x))\chi_U^i(x)$ and assume that x_0 is a point of density 1 for some U^{i_0} and show that there is no point of density 1 for U^k if $i_0 \ne k$.

Let $D_s(x_0)$ be a small square on which $\det Du$ has small *BMO* seminorm, while $\alpha^{i_0} = \operatorname{meas}(U^{i_0}\cap D_s(x_0))/\operatorname{meas}(D_s(x_0)) > 3/4$, so that $\alpha^k \le 1/4$ on $D_s(x_0)$. Since in general we can show that $\alpha^k(1-\alpha^k) \le C\epsilon$ on small squares inside Ω (see (3.18)) due to the *BMO* estimate, we may separate the values of α_k on small squares as either $\alpha^k \ge 1/2 + \mu_0$ or $\alpha_k \le 1/2 - \mu_0$ for some small $\mu_0 > 0$. Instead of "maximizing" the left-hand side of the above inequality (see (3.18) below), as in the proof of Lemma 1, we claim that there is no point of density 1 for U^k in $D_s(x_0)$. Otherwise, a continuous deformation of squares in $D_s(x_0)$ and the intermediate value theorem will lead to a square with $\alpha^k = 1/2$, which contradicts the estimates for α^k. Then a continuation argument will finish the induction process.

Proof of Lemma 2. We use the notation in the proof of Lemma 1 except for the subscript j. If $k = 1$, we have nothing to prove. If for $k-1 \ge 1$ the statement is true, we consider the case k. Suppose $Du(x) \in K_\epsilon$; we also have $|P_{E_{\bar\partial}}(Du(x)| \le \epsilon$ a.e. in Ω. Let $F(x) = P_{E_{\bar\partial}}(Du(x))$ for $x \in \Omega$; then u is a weak solution of

$$\Delta u = 2\operatorname{div}(P_{E_{\bar\partial}}(Du(x))) = 2\operatorname{div} F(x).$$

From Proposition 3, we have, for each $x_0 \in \Omega$, $0 < \rho < R$ such that $\bar B_\rho(x_0) \subset$

$\bar{B}_R(x_0) \subset \Omega$, and there is some $0 < \gamma < 2$ such that

(3.15)
$$\int_{B_\rho(x_0)} |Du - [Du]_{B_\rho(x_0)}|^2 dx \le C_1 \left(\frac{\rho}{R}\right)^{\gamma+2} \int_{B_R(x_0)} |Du - [Du]_{B_R(x_0)}|^2 dx + C_1 \epsilon^2 \rho^2,$$

where $C_1 > 0$ is a constant. If we choose $R > 0$ such that $\bar{B}_{2R}(x_0) \subset \Omega$, then for any $0 < \rho < R$ in (3.15),

(3.16)
$$\begin{aligned} &⨍_{B_\rho(x_0)} |Du - [Du]_{B_\rho(x_0)}|^2 dx \le C_2 \left(\frac{\rho}{R}\right)^{\gamma} ⨍_{B_R(x_0)} |Du - [Du]_{B_R(x_0)}|^2 dx + C_2 \epsilon^2 \\ &\le 16 C_2 \lambda_k^2 \left(\frac{\rho}{R}\right)^{\gamma} + C_2 \epsilon^2. \end{aligned}$$

Here we have used the fact that $\|Du\|_{L^\infty(\Omega, M^{2\times 2})} \le 2(\lambda_k + \epsilon) \le 4\lambda_k$ if we require $\epsilon < \lambda_k$. Thus if we choose $0 < \rho < R$ such that $\rho^\gamma < R^\gamma \epsilon^2 / C_2 (4\lambda_k)^2$, then $\|Du\|_{BMO(D_\rho(x_0))} \le C_3 \epsilon$ for squares centered at x_0 with side-length ρ, where $C_3 > 0$ is a constant depending on λ_k. Thus for every $x_0 \in \Omega$, there is $r = r(x_0, \Omega) > 0$ such that

$$\|Du\|_{BMO(D_r(x))} \le C_3 \epsilon. \tag{3.17}$$

Now since $Du(x) \in K_\epsilon$ in Ω, we write

$$Du(x) = \sum_{i=1}^{k} R(x)(\lambda_i I + f^i(x)) \chi_U^i(x), \tag{3.18}$$

where $\chi_{U^i}(x)$ is the characteristic function of $U^i = \{x \in \Omega,\, Du(x) \in (\lambda_i SO(2))_\epsilon\}$, while

$$\left\| \sum_{i=1}^{k} f^i \chi_{U^i} \right\|_{L^\infty(\Omega, M^{2\times 2})} \le \epsilon.$$

Since $\mathrm{meas}(\Omega \setminus (\cup_{i=1}^k (U_i)_\epsilon) = 0$ and U^i's are disjoint, without loss of generality, we may assume that $x_0 \in U^{i_0} \subset \Omega$ is a point of density 1. By definition of the density, there is a square $D_s(x_0) \subset D_r(x_0)$ with $s = s(x_0, \Omega)$ such that

$$\frac{\mathrm{meas}(U^{i_0} \cap D_s(x_0))}{\mathrm{meas}(D_s(x_0))} > \frac{3}{4}; \quad \text{hence} \quad \frac{\mathrm{meas}((\cup_{i\neq i_0}^k U^i) \cap D_s(x_0))}{\mathrm{meas}(D_s(x_0))} < \frac{1}{4}. \tag{3.19}$$

Notice that λ_1 and λ_k are two end points of the set $\{\lambda_i,\, 1 \le i \le k\}$. Now we claim that (i) if $\lambda_{i_0} \neq \lambda_k$, $\mathrm{meas}(U^k \cap D_s(x_0)) = 0$; (ii) if $\lambda_{i_0} \neq \lambda_1$, $\mathrm{meas}(U^1 \cap D_s(x_0)) = 0$. Supposing the assumption of (i) is satisfied, we use the Jacobian again as in the proof of Lemma 1. Since (3.17) and the boundedness of Du in L^∞ imply

$$⨍_Q |\det(Du) - [\det(Du)]_Q| dx \le C\epsilon \tag{3.20}$$

for every square $Q \subset D_r(x_0)$, where $C > 0$ depends on λ_k, we substitute (3.18) into (3.20) to obtain

$$⨍_Q \left| \sum_{i=1}^{k} \lambda_i^2 \chi_{U^i} - \left[\sum_{j=1}^{k} \lambda_j^2 \chi_{U^j} \right]_Q \right| dx \le C_4 \epsilon,$$

where $C_4 > 0$ depends on λ_k. Letting $\alpha^i = \operatorname{meas}(Q \cap U^i)/\operatorname{meas}(Q)$, $i = 1, \dots, k$, we have $\sum_{i=1}^k \alpha^i = 1$, $\alpha_i \geq 0$, and

$$\fint_Q \left| \sum_{i=1}^k \lambda_i^2 \chi_{U^i} - \left[\sum_{j=1}^k \lambda_j^2 \chi_{U^j} \right]_Q \right| dx \geq \fint_Q \left| \lambda_k^2 \chi_{U^k} - \left[\sum_{j=1}^k \lambda_j^2 \chi_{U^j} \right]_Q \right| \chi_{U^k} dx$$

$$= \alpha_k \left| \sum_{i=1}^k (\lambda_k^2 - \lambda_i^2)\alpha^i \right| \geq \alpha_k \sum_{i=1}^{k-1} (\lambda_k^2 - \lambda_{k-1}^2)\alpha^i = (\lambda_k^2 - \lambda_{k-1}^2)\alpha_k(1-\alpha_k) \geq r_K \alpha_k(1-\alpha_k).$$

Combining the last two sets of inequalities, we have

$$\alpha_k(1-\alpha_k) \leq C_4 \epsilon / r_K$$

on any square $Q \subset D_r(x_0)$. Note that the inequality above holds for any square $Q \subset D_r(y)$, $y \in \Omega$ and $r = r(y, \Omega)$. Now we require that $\epsilon < r_K/(4C_4)$. Then

$$\frac{\operatorname{meas}(U^k \cap Q)}{\operatorname{meas}(Q)} \left(1 - \frac{\operatorname{meas}(U^k \cap Q)}{\operatorname{meas}(Q)} \right) \leq \frac{1}{4} - \mu_0 \tag{3.21}$$

for some $0 < \mu_0 < 1/4$. Solving inequality (3.21), we have either

$$\text{(a)} \quad \frac{\operatorname{meas}(U^k \cap Q)}{\operatorname{meas}(Q)} \leq \frac{1}{2} - \sqrt{\mu_0}, \quad \text{or} \quad \text{(b)} \quad \frac{\operatorname{meas}(U^k \cap Q)}{\operatorname{meas}(Q)} \geq \frac{1}{2} + \sqrt{\mu_0} \tag{3.22}$$

for all squares $Q \subset D_r(x_0)$. We claim that (3.22(b)) cannot happen if $Q \subset D_s(x_0)$. If the claim is not true, there is some $Q_0 \subset D_s(x_0)$ such that (3.22(b)) holds. Since on $Q = D_s(x_0)$, (3.19) holds, and hence (3.22(a)) must hold, we may find a continuous family of squares $Q(t) \subset D_s(x_0)$, $0 \leq t \leq 1$, with $Q(0) = D_s(x_0)$ and $Q(1) = Q_0$, and show that there is some $Q(t_0) \subset D_s(x_0)$, $0 < t_0 < 1$, such that

$$\frac{\operatorname{meas}(U^k \cap Q(t_0))}{\operatorname{meas}(Q(t_0))} = \frac{1}{2}.$$

However, such a $Q(t_0)$ satisfies neither (3.22(a)) nor (3.22(b)), a contradiction.

Since (3.22(a)) holds for every nondegenerate square in $D_s(x_0)$, we may conclude that the points of density 1 for U^k in $D_s(x_0)$ are an empty set, and hence $\operatorname{meas}(U^k \cap D_s(x_0)) = 0$. Consequently, $Du(x) \in \cup_{i=1}^{k-1} (\lambda_i SO(2))_\epsilon$ a.e. in $D_s(x_0)$.

Now we show that $Du(x) \in \cup_{i=1}^{k-1} (\lambda_i SO(2))_\epsilon$ a.e. in Ω. Since Ω is arcwise connected, for any $x_1 \in \Omega$, there is a piecewise affine curve $\gamma : [0,1] \to \Omega$ connecting x_0 and x_1, with each piece of affine subcurve parallel to one of the coordinate axes such that $\operatorname{dist}(\gamma, \partial\Omega) = R_0 > 0$. So we may assume that there is some $0 < r_0 < R$ such that (3.17) holds on $D_{r_0}(x)$ for every $x \in \gamma$ and either (3.22(a)) or (3.22(b)) holds (note that (3.22(a)) and (3.22(b)) are independent of Q_0). Starting from $x_0 = \gamma(0)$ for sufficiently small $s > 0$, we consider a family of squares $Q(t) = D_s(\gamma(t))$ and a continuous function

$$f_s(t) = \frac{\operatorname{meas}(U^k \cap Q(t))}{\operatorname{meas}(Q(t))}.$$

Since $f_s(0) = 0$ as we proved earlier and either $f_s(t) \leq 1/2 - \sqrt{\mu_0}$ or $f_s(t) \geq 1/2 + \sqrt{\mu_0}$, we see that $f_s(t) \leq 1/2 - \sqrt{\mu_0}$ for all $t \in [0, 1]$. Hence $f_s(1) \leq 1/2 - \sqrt{\mu_0}$. Equivalently, we have

$$\frac{\operatorname{meas}(U^k \cap D_s(x_1))}{\operatorname{meas}(D_s(x_1))} \leq \frac{1}{2} - \sqrt{\mu_0}$$

for sufficiently small $s > 0$. Thus x_1 and, in general, every point in Ω are not points of density 1 for U^k; hence $\text{meas}(U^k) = 0$. Therefore $Du(x) \in \cup_{i=1}^{k-1}(\lambda_i SO(2))_\epsilon$ a.e. in Ω. The conclusion then follows from the induction assumption.

If the assumption of (ii) is satisfied, that is, $\lambda_{i_0} \neq \lambda_1$, a similar argument will lead to the conclusion that $\text{meas}(U^1) = 0$. Again by induction assumption, $Du(x) \in (\lambda_{i_0} SO(2))_\epsilon$ a.e. in Ω. □

Proof of Theorem 1. Let $\epsilon > 0$ satisfy the requirements of Theorem 2 and Lemmas 1 and 2. Since $\lim_{j\to 0} \int_\Omega \text{dist}(Du_j\, K_\epsilon)dx = 0$, there is a subsequence $u_{j_s} \rightharpoonup u$ in $W^{1,1}(\Omega, \mathbb{R}^2)$ with $\{\nu_x\}_{x\in\Omega}$ the corresponding gradient Young measures. This is because K_ϵ is compact, and hence $|Du_j|$ is equi-integrable in Ω. Therefore,

$$\lim_{s\to\infty} \int_\Omega \text{dist}(Du_{j_s}\, K_\epsilon)dx = \int_\Omega \int_{K_\epsilon} \text{dist}(\tau,\, K_\epsilon)d\nu_x(\tau)\, dx = 0,$$

which implies that $\text{supp}\,\nu_x \subset K_\epsilon$ a.e. Then for almost every $x_0 \in \Omega$, we see, from Proposition 2, that there is a homogeneous Young measure $\mu_x = \nu = \nu_{x_0}$, $x \in D$ a.e., supported in K_ϵ with the generating sequence $Du(x_0) + D\phi_{j_s}$ (up to yet another subsequence; see Proposition 4) with $\phi_{j_s} \in W_0^{1,\infty}(D, \mathbb{R}^2)$, and ϕ_{j_s} converges in the weak-$*$ sense to 0 in $W_0^{1,\infty}(D, \mathbb{R}^2)$, where D is the unit square in $\mathbb{R}^2$. Let $A = Du(x_0)$. Since $\lim_{k\to\infty} \int_D \text{dist}^p(A + D\phi_{j_s}(x), K_\epsilon)dx = 0$ for all $1 \leq p < \infty$, we may apply Theorem 2 to the quasi-convex function $F_\epsilon(\cdot)$ given by Theorem 2 to conclude that $Du(x_0) = A \in K_\epsilon$, and hence $Du(x_0) \in (\lambda_{i_0} SO(2))_\epsilon$ for some $1 \leq i_0 \leq k$. Lemma 1 then implies that

$$\lim_{k\to\infty} \int_D \text{dist}^p(A + D\phi_{j_k}(x),\, (\lambda_{i_0} SO(2))_\epsilon)dx = 0.$$

Thus $\text{supp}\,\nu_{x_0} = \text{supp}\,\nu \subset (\lambda_{i_0} SO(2))_\epsilon$ and $Du(x_0) \in (\lambda_{i_0} SO(2))_\epsilon$. Since this is true for almost every $x_0 \in \Omega$, we may claim that $\text{supp}\,\nu_x \subset (\lambda_{i_x} SO(2))_\epsilon$ if $Du(x) \in (\lambda_{i_x} SO(2))_\epsilon$ for some $1 \leq i_x \leq k$.

Since in any case we have $Du(x) \in K_\epsilon$ a.e., from Lemma 2 we see that $Du(x) \in (\lambda_{i_0} SO(2))_\epsilon$ a.e. in Ω for some fixed $1 \leq i_0 \leq k$. Combining this with the last paragraph, we see that $\text{supp}\,\nu_x \subset (\lambda_{i_0} SO(2))_\epsilon$ a.e. in Ω, that is,

$$\lim_{k\to\infty} \int_D \text{dist}(Du_{j_s}(x),\, (\lambda_{i_0} SO(2))_\epsilon)dx = 0. \qquad \square$$

4. Proof of Theorem 2. The quasi-convex relaxation for $Q\,\text{dist}^2(A,\, SO(2)H)$ was obtained in [13]. We first show that

(4.1)
$$Q\,\text{dist}^2(A, K) = C_{E_\partial}[\text{dist}^2(P_{E_\partial}(A), K) + |P_{E_\partial}(A)|^2] + [|P_{E_{\bar\partial}}(A)|^2 - |P_{E_\partial}(A)|^2].$$

Clearly

$$\begin{aligned} F(A) &:= \text{dist}^2(A, K) \\ &= [\text{dist}^2(P_{E_\partial}(A), K) + |P_{E_\partial}(A)|^2] + [|P_{E_{\bar\partial}}(A)|^2 - |P_{E_\partial}(A)|^2] \\ &= V(A) + H(A) \qquad (4.2) \\ &\geq C_{E_\partial}[\text{dist}^2(P_{E_\partial}(A), K) + |P_{E_\partial}(A)|^2] + [|P_{E_{\bar\partial}}(A)|^2 - |P_{E_\partial}(A)|^2] \\ &= G(A) + H(A), \end{aligned}$$

where $G(A) = C_{E_\partial} V(A)$, and the right-hand side of (4.2) is quasi-convex because G is convex and $2H(A) = -\det A$ is quasi-convex. Thus

$$Q\,\mathrm{dist}^2(A, K) \geq G(A) + H(A). \tag{4.3}$$

Next we show that $R\,\mathrm{dist}^2(A, K) \leq G(A) + H(A)$. To see this we use the Kohn–Strang scheme (2.1). Let $\Lambda = \{a \otimes b \in M^{2\times 2},\, a,\, b \in \mathbb{R}^2\}$ be the set of all rank-one matrices in $M^{2\times 2}$. It is easy to check that the mapping $P_{E_\partial} : \Lambda \to E_\partial$ is an onto map. Also $H(a \otimes b) = 0$ for every $a \otimes b \in \Lambda$. We let $P_{E_\partial}(A) = X$, Y_1, Y_2 be arbitrary matrices in E_∂ satisfying $\lambda Y_1 + (1-\lambda)Y_2 = X$, and we let $Z = (Y_1 - Y_2)$. Since $P_{E_\partial} : \Lambda \to E_\partial$ is onto, there is a rank-one matrix $B \in \Lambda$ such that $P_{E_\partial}(B) = Z$. Now we estimate $R_1 F(A)$ by lifting the matrices in E_∂ to $M^{2\times 2}$. We have $Y_1 = X + (1-\lambda)Z$, $Y_2 = X - \lambda Z$. We let $A_1 = A + (1-\lambda)B$, $A_2 = A - \lambda B$; then

$$\lambda A_1 + (1-\lambda)A_2 = A, \quad A_1 - A_2 = B \in \Lambda, \quad P_{E_\partial}(A_1) = Y_1, \quad P_{E_\partial}(A_2) = Y_2.$$

Thus

$$\begin{aligned} R_1 F(A) &\leq \lambda F(A_1) + (1-\lambda) F(A_2) \\ &= [\lambda V(P_{E_\partial}(A_1)) + (1-\lambda) G(P_{E_\partial}(A_2))] + [\lambda H(A_1) + (1-\lambda) H(A_2)] \\ &= I_1 + I_2. \end{aligned}$$

We have

$$\begin{aligned} I_1 &= \lambda V(P_{E_\partial}(A_1)) + (1-\lambda) V(P_{E_\partial}(A_2)) = \lambda V(Y_1) + (1-\lambda) V(Y_2), \\ I_2 &= \lambda H(A_1) + (1-\lambda) H(A_2) = \lambda H(A + (1-\lambda)B) + (1-\lambda) H(A - \lambda B). \end{aligned}$$

Since for a quadratic form, we have $H(A + P) = H(A) + H(B) + DH(A)B$, where $DH(A)$ is the gradient of H at A, which is linear in A, then

$$\begin{aligned} &\lambda H(A + (1-\lambda)B) = \lambda H(A) + \lambda H((1-\lambda)B) + \lambda(1-\lambda) DH(A)B \\ &= \lambda H(A) + \lambda(1-\lambda) DH(A)B. \end{aligned}$$

Here we have used the fact $H(B) = 0$. Similarly, we have

$$(1-\lambda) H(A - \lambda B) = (1-\lambda) H(A) - (1-\lambda)\lambda DH(A)B.$$

Hence $I_2 = \lambda H(A) + \lambda(1-\lambda) DH(A)B + (1-\lambda) H(A) - (1-\lambda)\lambda DH(A)B = H(A)$. Consequently, we obtain $R_1 F(A) \leq [\lambda V(Y_1) + (1-\lambda) V(Y_2)] + H(A)$. Taking infimum on Y_1, Y_2 with $\lambda Y_1 + (1-\lambda)Y_2 = X$ (cf. (2.2)), we obtain

$$R_1 F(A) \leq C_1 V(A) + H(A). \tag{4.4}$$

Repeating the previous step by using (4.4), we see that $R_k F(A) \leq C_k V(A) + H(A)$. Passing to the limit $k \to \infty$, we have $RF(A) \leq [C_{E_\partial} V(P_{E_\partial}(A))] + H(A)$. So the reversed inequality of (4.3) is reached. Since we also have $RF(A) \geq QF(A) \geq C_{E_\partial} V(P_{E_\partial}(A)) + H(A)$, the proof of $Q\,\mathrm{dist}^2(A, K) = G(A) + H(A)$ is complete.

The estimates in (i) can be easily deduced from the following lemma, which was established in [31] by taking $\lambda = 1$.

LEMMA 3. *Let $K \subset \mathbb{R}^n$ be a closed set with $K \neq \mathbb{R}^n$. Then*

$$C[\mathrm{dist}^2(x, K) + \lambda|x|^2] - \lambda|x|^2 \geq \frac{\lambda}{1+\lambda}\,\mathrm{dist}^2(x, K)$$

for $x \in \mathbb{R}^n$, where $\lambda > 0$ is a constant.

Notice that if $K = \mathbb{R}^n$, the above inequality is trivially true.

Now we prove (ii). Note that (iii) is a direct consequence of (ii). We may view the problem in (ii) as a problem in the Euclidean space $\mathbb{R}^2$: Suppose $K \subset \mathbb{R}^2$ consists of finitely many circles, in our case, $K = \cup_{i=1}^k S_{\sqrt{2}\lambda_i} \subset \mathbb{R}$, where $S_\lambda = \{x \in \mathbb{R}^2,\ |x| = \lambda\}$. Let $\hat{K} = \{\sqrt{2}\lambda_i\}_{i=1}^k$. Then squared distance function can be written as $\operatorname{dist}^2(x, K) = \operatorname{dist}^2(|x|, \hat{K})$. We may view the function $[\operatorname{dist}^2(|x|, \hat{K}) + \lambda|x|^2]$ as an even function of one variable $\operatorname{dist}^2(t, \hat{K} \cup -\hat{K}) + \lambda t^2 := h(t)$, where $-\hat{K} = \{-\sqrt{2}\lambda_i\}_{i=1}^k$. We show that when t is close to $\hat{K} \cup -\hat{K}$, $C[h(t)] = h(t)$.

We give an estimate for the general case when $L \subset \mathbb{R}$ is any finite set and $f(x) = \operatorname{dist}^2(x, L) + \lambda|x|^2$ for $x \in \mathbb{R}$. Since for a function f defined on $\mathbb{R}$, we have (see [23])

$$Cf(x) = \sup\{l(x),\ l(y) \le f(y),\ y \in \mathbb{R},\ l \text{ affine}\}.$$

It suffices if we can show for y_0 near L and $\operatorname{dist}(y_0, L) = |y_0 - x_0| < \epsilon$ for some $\epsilon > 0$ to be determined, $x_0 \in L$, that the tangent line of the graph of f at $(y_0, f(y_0))$ stays underneath the graph of f. The tangent line at y_0 is given by

$$l(y) = (y_0 - x_0)^2 + \lambda y_0^2 + 2(y - x_0)(y - y_0) + 2\lambda y_0(y - y_0).$$

Let us estimate $f(y)$. If $\operatorname{dist}(y, L) = |y - x_0|$, then $f(y) \ge l(y)$ because of the convexity of f near x_0. Let $d = \min\{|x - y|,\ x \ne y,\ x,\ y \in L\} > 0$. If $\operatorname{dist}(y, L) = |y - x_1| < |y - x_0|$ for some $x_1 \in L$, we have

$$\begin{aligned} f(y) &= (y - x_1)^2 + \lambda y^2 = (y - x_0)^2 + \lambda y^2 + 2(y - x_0)(x_0 - x_1) + (x_0 - x_1)^2 \\ &= G(y) + H(y), \end{aligned}$$

where $G(y) = (y - x_0)^2 + \lambda y^2$ and $H(y) = 2(y - x_0)(x_0 - x_1) + (x_0 - x_1)^2$. We then have, by Taylor's expansion of G at y_0, that $G(y) = l(y) + (1 + \lambda)(y - y_0)^2$, and hence

$$\begin{aligned} &f(y) - l(y) = (1 + \lambda)(y - y_0)^2 + H(y) \\ &\ge (1 + \lambda)(y - y_0)^2 - 2|y - x_0||x_1 - x_0| + (x_0 - x_1)^2 \\ &\ge [(1 + \lambda)(y - y_0)^2 - 2|y - y_0||x_1 - x_0| + (x_0 - x_1)^2] - 2|x_0 - y_0||x_1 - x_0| \\ &\ge \frac{\lambda}{1 + \lambda}(x_0 - x_1)^2 - 2|x_0 - y_0||x_1 - x_0| \\ &\ge |x_0 - x_1| \left(\frac{\lambda}{1 + \lambda} d - |x_0 - y_0| \right) \ge 0 \end{aligned}$$

whenever $|y_0 - x_0| \le \lambda d/(1 + \lambda)$. Therefore we have $Cf(y_0) = f(y_0)$ when $\operatorname{dist}(y_0, L) \le \lambda d/(1 + \lambda)$. Hence we conclude that $R \operatorname{dist}^2(A, K) = Q \operatorname{dist}^2(A, K) = \operatorname{dist}^2(A, K)$ whenever $\operatorname{dist}(A, K) \le \lambda d/(1 + \lambda)$. Setting $\lambda = 1$ in our case and calculating

$$\begin{aligned} d &= \min\{|X - Y|,\ X, Y \in P_{E\partial}(K),\ X \ne Y\} \\ &= \sqrt{2} \min\{\lambda_{i+1} - \lambda_i,\ i = 1, \dots, k - 1\} = \sqrt{2} g_K, \end{aligned}$$

thus $R \operatorname{dist}^2(A, K) = Q \operatorname{dist}^2(A, K) = \operatorname{dist}^2(A, K)$ whenever $\operatorname{dist}^2(A, K) \le \frac{\sqrt{2}}{2} g_K$.

Acknowledgment. I wish to thank the referees for their helpful suggestions.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] E. Acerbi and N. Fusco, *Semicontinuity problems in the calculus of variations*, Arch. Ration. Mech. Anal., 86 (1984), pp. 125–145.
[3] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Ration. Mech. Anal., 63 (1977), pp. 337–403.
[4] J. M. Ball, *A version of the fundamental theorem of Young measures*, in Partial Differential Equations and Continuum Models of Phase Transitions, M. Rascle, D. Serre, and M. Slemrod, eds., Lecture Notes in Phys. 334, Springer-Verlag, Berlin, 1989, pp. 207–215.
[5] K. Bhattacharya, N. B. Firoozye, R. D. James, and R. V. Kohn, *Restrictions on Microstructures*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 843–878.
[6] B. Bojarski and T. Iwaniec, *Quasiconformal mappings and non-linear elliptic equations in two variables* I, Bull. Acad. Polon. Sci. Sèr. Math. Astronom. Phys., 22 (1974), pp. 473–478.
[7] B. Bojarski and T. Iwaniec, *Quasiconformal mappings and non-linear elliptic equations in two variables* II, Bull. Acad. Polon. Sci. Sèr. Math. Astronom. Phys., 22 (1974), pp. 479–488.
[8] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Ration. Mech. Anal., 100 (1987), pp. 13–52.
[9] J. M. Ball and R. D. James, *Proposed experimental tests of a theory of fine microstructures and the two-well problem*, Philos. Trans. Roy. Soc. London Sect. A, 338 (1992), pp. 389–450.
[10] J. M. Ball and R. D. James, *Personal communication*, 1993.
[11] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.
[12] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[13] N. B. Firoozye, *Optimal use of the translation method and relaxations of variational problems*, Comm. Pure Appl. Math., 44 (1991), pp. 643–678.
[14] M. Giaquinta, *Introduction to Regularity Theory for Nonlinear Elliptic Systems*, Birkhäuser-Verlag, Basel, 1993.
[15] D. Kinderlehrer and P. Pedregal, *Characterizations of Young measures generated by gradients*, Arch. Ration. Mech. Anal, 115 (1991), pp. 329–365.
[16] R. V. Kohn and D. Strang, *Optimal design and relaxation of variational problems* I, Comm. Pure Appl. Math., 39 (1986), pp. 113–137.
[17] R. V. Kohn and D. Strang, *Optimal design and relaxation of variational problems* II, Comm. Pure Appl. Math., 39 (1986), pp. 139–182.
[18] R. V. Kohn and D. Strang, *Optimal design and relaxation of variational problems* III, Comm. Pure Appl. Math., 39 (1986), pp. 353–377.
[19] J. P. Matos, *Young measures and the absence of fine microstructures in a class of phase transitions*, European J. Appl. Math., 3 (1992), pp. 31–54.
[20] C. B. Morrey, Jr., *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, 1966.
[21] S. Müller, *Variational models for microstructure and phase transitions*, in Calculus of Variations and Geometric Evolution Problems, Lecture Notes in Math. 1713, Springer-Verlag, Berlin, pp. 85–210.
[22] S. Müller, *A sharp version of Zhang's theorem on truncating sequences of gradients*, Trans. Amer. Math. Soc., 351 (1999), pp. 4585–4597.
[23] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[24] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[25] V. Šverák, *On Regularity of the Monge-Ampère Equation without Convexity Assumptions*, preprint, 1990.
[26] V. Šverák, *Quasiconvex functions with subquadratic growth*, Proc. Roy. Soc. London Ser. A, 433 (1991), pp. 723–725.
[27] V. Šverák, *Rank one convexity does not imply quasiconvexity*, Proc. Roy. Soc. Edinburgh Sect. A, 120 (1992), pp. 185–189.
[28] V. Šverák, *On the problem of two wells*, in Microstructure and Phase Transitions, IMA Vol. Math. Appl. 54, Springer-Verlag, New York, 1993, pp. 183–189.
[29] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics, Heriot-Watt Symposium IV, R. J. Knops, ed., Pitman, Boston, MA, 1979.

[30] K.-W. Zhang, *A construction of quasiconvex functions with linear growth at infinity*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 19 (1992), pp. 313–326.
[31] K.-W. Zhang, *Maximal extension for linear spaces of real matrices with large rank*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 1481–1491.
[32] K.-W. Zhang, *On separation of gradient Young measures*, in Calc. Var. Partial Differential Equations, to appear.

SIAM J. MATH. ANAL.
Vol. 34, No. 5, pp. 1226–1251

L^1 STABILITY FOR SYSTEMS OF HYPERBOLIC CONSERVATION LAWS WITH A RESONANT MOVING SOURCE*

SEUNG-YEAL HA† AND TONG YANG‡

Abstract. In this paper, we study the L^1 stability for the system $u_t + f(u)_x = g(x-ct,u)$ when one of the characteristic fields has resonance with the moving source. The nonlinear resonance occurs when the speed of the source can coincide with one of the characteristic speeds of the hyperbolic conservation laws. In this situation, a wave pattern can be either stable or unstable. By employing a nonlinear functional approach, we prove the L^1 stability of a transonic shock wave under the stability conditions introduced in [W.-C. Lien, *Comm. Pure Appl. Math.,* 52 (1999), pp. 1075–1098; T.-P. Liu, *Comm. Math. Phys.,* 83 (1982), pp. 243–260].

Key words. L^1 nonlinear functional, hyperbolic conservation laws, resonant source

AMS subject classifications. 35L65, 35L67

PII. S0036141001397983

1. Introduction.

The purpose of this paper is to study the L^1 stability for systems of hyperbolic conservation laws with a resonant moving source:

$$\begin{cases} u_t + f(u)_x = g(x-ct,u), & (x,t) \in \mathbb{R} \times \mathbb{R}_+, \\ u(x,0) = u_0(x), & x \in \mathbb{R}, \end{cases} \tag{1.1}$$

where $u \in \mathcal{N} \subset \mathbb{R}^n$, $f : \mathcal{N} \to \mathbb{R}^n$, and $g : \mathbb{R} \times \mathcal{N} \to \mathbb{R}^n$ denote the conserved quantities, the C^3 flux function, and the source, respectively. Here $\mathcal{N}$ is a small neighborhood of some reference state u_0 in $\mathbb{R}^n$. This system is assumed to be strictly hyperbolic with the ith characteristic field being resonant with the source, i.e., $\lambda_i(u) \approx c$. It is well known [13] that in general, the system (1.1) does not admit a classical solution even for the smooth initial data because of the nonlinearity of the flux function. Therefore, one needs to consider weak solutions.

DEFINITION 1.1. *A bounded measurable function $u(x,t)$ is a weak solution of the system* (1.1) *with given initial data $u_0(x)$ if and only if*

$$\int_0^\infty \int_{-\infty}^\infty [u\phi_t + f(u)\phi_x + g(x-ct,u)\phi](x,t)dxdt + \int_{-\infty}^\infty u_0(x)\phi(x,0)dx = 0$$

$$\textit{for any } \phi \in C_c^1(\mathbb{R}^2).$$

Hyperbolic systems of conservation laws with sources appear in many physical situations, such as three-dimensional compressible Euler equations with symmetries

*Received by the editors November 13, 2001; accepted for publication (in revised form) September 6, 2002; published electronically April 30, 2003.
http://www.siam.org/journals/sima/34-5/39798.html
†Department of Mathematics, University of Wisconsin–Madison, 480 Lincoln Drive, Madison, WI 53706 (ha@math.wisc.edu).
‡Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong (matyang@sobolev.cityu.edu.hk). The research of this author was partially supported by the RGC Competetive Earmarked Research grant of Hong Kong 9040468.

[6], a flow through a duct of variable cross section [6, 14, 19, 18, 17, 16, 24], and a moving magnetic field for magneto-hydrodynamics (MHD) [12]. A prototype of these systems is a quasi-one-dimensional nozzle flow model:

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial(\rho u)}{\partial x} = -\dfrac{A'(x)}{A(x)}(\rho u), \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+, \\ \dfrac{\partial(\rho u)}{\partial t} + \dfrac{\partial(\rho u^2 + P)}{\partial x} = -\dfrac{A'(x)}{A(x)}(\rho u^2), \\ \dfrac{\partial(\rho E)}{\partial t} + \dfrac{\partial(\rho u E + P u)}{\partial x} = -\dfrac{A'(x)}{A(x)}(\rho u E + P u), \\ P = P(e, \rho), \end{cases}$$

where $A(x)$ is the cross sectional area of a nozzle, ρ is the density, u is the velocity, P is the pressure, e is the internal energy, and $E = e + \frac{u^2}{2}$ is the total energy of a gas.

The global existence and time-asymptotic stability of (1.1) were studied in [14, 19, 18, 17, 16]. Recently in [2, 11], the L^1 stability of small initial data was studied for systems with a nonresonant moving source, i.e., in the case in which the characteristic speeds are strictly different from that of the source. In contrast, for the system with a resonant moving source, the global existence of weak solutions with bounded total variation (BV) has been studied only for special initial data such as a small perturbation of one transonic shock wave moving with speed c in [14, 18, 17], and its time-asymptotic stability and instability were shown to be strongly dependent on the source. Moreover, the time-asymptotic stability of a steady transonic shock wave under the stability condition (1.3) or (1.4) was analyzed in [14, 18]. Under the condition (1.3), a simple wave pattern consisting of one steady transonic shock wave in a nozzle is stable time-asymptotically so that the perturbed transonic shock wave is still observed in the nozzle. However, under the condition (1.4), the perturbed transonic shock wave is unstable time-asymptotically so that it is not observed in a nozzle after a finite time [9, 19, 18, 17]. Let $u_0 \in \mathcal{N}$ be a fixed reference state, and assume that the right (left) eigenvectors $r_i(u_0)$, $l_j(u_0)$ form dual bases such that

$$l_i(u_0) \cdot r_j(u_0) = \delta_{ij}, \quad 1 \le i,\ j \le n.$$

And by using a moving coordinate with speed c, we may assume that c equals zero; i.e.,

$$u_t + f(u)_x = g(x,u). \tag{1.2}$$

Set

$$u = \sum u^i r_i(u_0), \quad f(u) = \sum f^i(u) r_i(u_0), \quad g(x,u) = \sum g^i(x,u) r_i(u_0).$$

Then it follows from (1.2) that

$$\begin{cases} u^i = l_i(u_0) \cdot u, \quad f^i(u) = l_i(u_0) \cdot f(u), \quad g^i(x,u) = l_i(u_0) \cdot g(x,u), \\ u^i_t + f^i(u)_x = g^i(x,u), \quad 1 \le i \le n. \end{cases}$$

In the following, we assume that the source $g(x,u)$ has support $[0,1]$ in x and define a function $G(x)$ as follows.

$$g(x,u)=0, \quad x\notin(0,1), \quad G(x)\equiv \sup_{p\in\mathcal{N}}\left\{|g(x,p)|+\left\|\frac{\partial g(x,p)}{\partial p}\right\|\right\}.$$

For the resonant ith characteristic field, the following time-asymptotic stability and instability conditions were introduced in [14, 18]:

$$G_{ii}(x,u)\doteq l_i(u_0)\frac{\partial g(x,u)}{\partial u}r_i(u_0)<0 \quad \text{(stability)}, \tag{1.3}$$

$$G_{ii}(x,u)\doteq l_i(u_0)\frac{\partial g(x,u)}{\partial u}r_i(u_0)>0 \quad \text{(instability)}, \tag{1.4}$$

where we have used the notation $G_{ij}(x,u)\doteq l_i(u_0)\frac{\partial g(x,u)}{\partial u}r_j(u_0)$. In this paper, for the sake of definiteness, we assume the first characteristic field is resonant with the source. In the study of L^1 stability for the system (1.1), in order to have a uniform L^1 stability in time, one needs to control the error caused by the source, in addition to the analysis of the homogeneous hyperbolic conservation laws. In [2, 11], when the source has no resonance with any of the characteristic fields, a quadratic functional $Q_{so}(t)$ was introduced to control these errors in L^1 analysis:

$$\begin{aligned}
&Q_{so}(t)=\sum_{j=1}^{n}Q_{so}^j(t),\\
&Q_{so}^j(t)\equiv\begin{cases}\int_{-\infty}^{\infty}|q_j(x,t)|\left(\int_{x(q_j)}^{\infty}G(x)dx\right) & \text{if } \lambda(q_j)>0,\\ \int_{-\infty}^{\infty}|q_j(x,t)|\left(\int_{-\infty}^{x(q_j)}G(x)dx\right) & \text{if } \lambda(q_j)<0.\end{cases}
\end{aligned}$$

As long as the characteristic speeds of the system (1.2) are strictly away from zero, $Q_{so}(t)$ gives a good decay term with respect to time which is enough to control the errors caused by the source. However, for the resonant field $(\lambda_1(u),r_1(u))$, since $\lambda_1(u)$ has a definite sign depending on the relative location compared to the relatively strong shock wave, we need to modify the $Q_{so}^1(t)$, defined as in [2, 11], so that it gives the decay estimate:

$$\begin{aligned}
Q_{so}^1(t)=&\int_{-\infty}^{x_f(t)}|q_1(x,t)|\left(\int_{x(q_1)}^{\infty}G(\xi)d\xi\right)dx\\
&+\int_{x_f(t)}^{\infty}|q_1(x,t)|\left(\int_{x_f(t)}^{x(q_1)}G(\xi)d\xi+\int_{x_f(t)}^{\infty}G(\xi)d\xi\right)dx,
\end{aligned}$$

where $x_f(t)$ is the location of the relatively strong shock at time t. In Lemma 3.3, we estimate the error of time variation of L^1 distance due to the source which is in the following form:

$$\begin{aligned}
\Gamma_{so}&\equiv\sum_{j}\sum_{\alpha^i\in J}\{\lambda(q_j^-(\alpha^i))|q_j^-(\alpha^i)|-\lambda(q_j^+(\alpha^i))|q_j^+(\alpha^i)|\}\\
&=\sum_{j=1}^{n}\mathcal{O}(1)\int_{-\infty}^{\infty}G(x)|q_j(x,t)|dx.
\end{aligned}$$

As shown in [2, 11], the quadratic functional $Q_{so}(t)$,

$$
\begin{aligned}
Q_{so}(t) &\equiv \sum_{j=1}^{n} Q_{so}^{j}(t),\\
Q_{so}^{1}(t) &= \int_{-\infty}^{x_f(t)} |q_1(x,t)| \left(\int_{x(q_1)}^{\infty} G(\xi) d\xi \right) dx\\
&\quad + \int_{x_f(t)}^{\infty} |q_1(x,t)| \left(\int_{x_f(t)}^{x(q_1)} G(\xi) d\xi + \int_{x_f(t)}^{\infty} G(\xi) d\xi \right) dx,\\
Q_{so}^{j}(t) &= \int_{-\infty}^{\infty} |q_j(x,t)| \left(\int_{x(q_j)}^{\infty} G(\xi) d\xi \right) dx, \quad j \geq 2,
\end{aligned}
$$

can compensate the Γ_{so} effectively, and a uniform L^1 stability follows. The main assumptions of this paper are as follows.

Main Assumptions.

1. The system (1.1) is strictly hyperbolic. Let $\lambda_i(u), (i \in \{1, \dots, n\})$ be distinct real eigenvalues of $f'(u)$, and let $r_i(u)$ ($l_i(u)$) be the corresponding right (left) eigenvectors of $f'(u)$; i.e.,

$$
\begin{aligned}
&f'(u) r_i(u) = \lambda_i(u) r_i(u), \qquad && \lambda_1(u) < \cdots < \lambda_n(u),\\
&l_i(u) f'(u) = \lambda_i(u) l_i(u), && l_i(u) \cdot r_j(u) = \delta_{ij}.
\end{aligned}
$$

2. Each characteristic field $(\lambda_j(u), r_j(u))$ is either genuinely nonlinear (g.n.l.) or linearly degenerate (l.d.g.) in the sense of Lax [13]:

$$
\begin{aligned}
&(\lambda_j(u), r_j(u)) \quad \text{is g.n.l.} \iff \nabla \lambda_j(u) \cdot r_j(u) \neq 0 \quad \text{for all } u \in \mathcal{N},\\
&(\lambda_j(u), r_j(u)) \quad \text{is l.d.g.} \iff \nabla \lambda_j(u) \cdot r_j(u) \equiv 0.
\end{aligned}
$$

3. The first characteristic field is g.n.l. and resonant with the source; that is,

$$
\nabla \lambda_1(u) \cdot r_1(u) \neq 0, \quad \lambda_1(u) \approx 0 \text{ for } u \in \mathcal{N}.
$$

4. $g(x,p)$ is piecewise differentiable in x, is continuously differentiable in p, has compact support in x, and is sufficiently weak in the following sense:

$$
\begin{cases}
g(x,p) = 0 \ \text{ if } \ x \notin [0,1],\\
G(x) \equiv \sup\limits_{p \in \mathcal{N}} \left\{ |g(x,p)| + \left\| \dfrac{\partial g(x,p)}{\partial p} \right\| \right\}, \quad G_1 \equiv \|G(\cdot)\|_{L^1(\mathbb{R})},\\
G_0 \equiv \|G(\cdot)\|_{L^\infty(\mathbb{R})}, \quad G_0 + G_1 \ll 1.
\end{cases}
$$

We notice that the case in which the source is not of compact support but has the above smallness properties can be treated similarly.

Without any smallness assumption on the source, we can expect only a local L^1 stability in time. That is, there exist a finite time T and constants $C, \bar{c} > 0$ such that

$$
||u(\cdot,t) - v(\cdot,t)||_{L^1(\mathbb{R})} \leq C e^{\bar{c}t} ||u(\cdot,0) - v(\cdot,0)||_{L^1(\mathbb{R})} \quad \text{when} \quad 0 < t < T.
$$

This local L^1 stability was first proved in [8]. For a "diagonally dominant dissipative

system," L^1 stability based on a nonlinear functional approach through the front tracking algorithm was studied in [1].

The main result of this paper is as follows. Let $u(x)$ be a stationary solution consisting of a transonic 1-shock wave connecting to a supersonic stationary wave $u_1(x)$ for $x < x_*$ and a subsonic stationary wave $u_2(x)$ for $x > x_*$ as follows:

$$u(x) = \begin{cases} u(-\infty), & x < 0, \\ u_1(x), & 0 \le x < x_*, \\ u_2(x), & x_* < x \le 1, \\ u(+\infty), & x > 1, \end{cases}$$

and $\lambda_1(u_2(x)) < 0 < \lambda_1(u_1(x))$, $|\lambda_1(u(x))| \ge \lambda_* > 0$, $\sigma(u_1(x_*-), u_2(x_*+)) = 0$. Here λ_* is a positive constant.

THEOREM 1.2. *Let $v(x,t)$ be a weak solution obtained by the Glimm scheme corresponding to initial data v_0, which is a small perturbation of $u(x)$ in* [14, 18]. *Assume that the condition* (1.3) *holds. Then the steady transonic* 1*-shock wave solution is uniformly L^1 stable, i.e.,*

$$||v(\cdot,t) - u(\cdot)||_{L^1(\mathbb{R})} \le G||v_0(\cdot) - u(\cdot)||_{L^1(\mathbb{R})},$$

where G is a generic constant independent of t,

Remark 1.1. Under the instability condition (1.4), it is well known [14, 18] that there exists a perturbation which can be arbitrarily small so that the perturbed big shock wave moves away from the nozzle and goes to infinity as time goes to infinity. Therefore, the L_1 distance between the stationary solution and the perturbed solution grows in time and can reach any given constant in finite time. The detailed analysis will show that the L_1 distance between these two solutions grows at least linearly in time in the general setting.

The rest of the paper is organized as follows. In section 2, we will first define a modified Glimm-type functional which is slightly different from the one in [14]. Then we will review the basic theory of the nonlinear functional approach for L^1 stability. In section 3, we will derive basic estimates on the shock strength and on how the source affects the L^1 distance. Finally, in section 4, by using the nonlinear functional defined in section 2, we will prove L^1 stability of a steady transonic 1-shock wave solution under the condition (1.3).

2. Preliminaries. In this section, we are going to define a modified Glimm-type functional, which is slightly different from the one in [14], and briefly present the simplified wave patterns given in [3, 11, 21, 22]. And then, we will define a nonlinear functional which is equivalent to the L^1 distance of two approximate weak solutions.

We first review the basics of the theory for the system of hyperbolic conservation laws:

$$\begin{aligned} &u_t + f(u)_x = 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+, \\ &u(x,0) = u_0(x), \quad x \in \mathbb{R}. \end{aligned} \tag{2.1}$$

A discontinuity (u_-, u_+) is called an i-shock wave for (2.1) with speed σ if it satisfies the Rankine–Hugoniot and entropy conditions

$$\begin{cases} f(u_+) - f(u_-) = \sigma(u_+ - u_-) \ \text{(R–H condition)}, \\ \lambda_i(u_+) < \sigma < \lambda_i(u_-) \ \text{(entropy condition)}. \end{cases}$$

The Riemann problem for (2.1) is the initial value problem with simple jump initial data

$$u(x,0) = \begin{cases} u_l, & x < 0, \\ u_r, & x > 0. \end{cases}$$

It is well known [13] that the Riemann solution is a function of $\frac{x}{t}$ and consists of $n+1$ constant states $\{u_l = u_0, u_1, u_2, \dots, u_n = u_r\}$, which are connected by shock waves, rarefaction waves, or contact discontinuities.

In the following, we define an ith rarefaction curve $R_i(u_0)$ and an ith shock curve $H_i(u_0)$.

$$\begin{aligned} R_i(u_0) \equiv & \text{ the integral curve of a vector field } r_i(u) \cdot \nabla_u \text{ passing through } u_0, \\ H_i(u_0) \equiv & \{u \in \mathbb{R}^n : \lambda_i(u_0,u)(u-u_0) = f(u) - f(u_0) \text{ for some scalar } \lambda_i(u_0,u), \\ & \max_u \lambda_{i-1}(u) < \lambda_i(u_0,u) < \min_u \lambda_{i+1}(u)\}. \end{aligned}$$

For an l.d.g. characteristic field $(\lambda_i(u), r_i(u))$, it is well known [13] that $H_i(u_0) = R_i(u_0)$. We parameterize these curves by the arc length ξ, and we divide the ith shock curve $H_i(u_0)$ and the ith rarefaction curve $R_i(u_0)$ as follows:

$$\begin{aligned} H_i^+(u_0) &\equiv \{u \in H_i(u_0) : \lambda_i(u) > \lambda_i(u_0,u) > \lambda_i(u_0)\}, \\ H_i^-(u_0) &\equiv \{u \in H_i(u_0) : \lambda_i(u) < \lambda_i(u_0,u) < \lambda_i(u_0)\}, \\ R_i^+(u_0) &\equiv \{u \in R_i(u_0) : \lambda_i(u) \geq \lambda_i(u_0)\}, \\ R_i^-(u_0) &\equiv \{u \in R_i(u_0) : \lambda_i(u) \leq \lambda_i(u_0)\}. \end{aligned}$$

Moreover, we define an ith wave curve as follows:

$$W_i(u_0) \equiv \begin{cases} H_i^-(u_0) \cup R_i^+(u_0) \text{ if } i\text{th characteristic field is g.n.l.}, \\ H_i(u_0) = R_i(u_0) \text{ if } i\text{th characteristic field is l.d.g.} \end{cases}$$

Then by the second order contact of the ith shock curve and the ith rarefaction curve at u_0, the ith wave curve $W_i(u_0)$ is a C^2-curve [13].

We consider a small perturbation of a steady transonic shock wave under the condition (1.3). In this case, the Glimm-type functional is defined as follows [10, 14]. Let $r = \Delta x$ and $s = \Delta t$ in the Glimm scheme. Let J be a space like curve in $x - t$ plane. We use x_J to denote the location of the big shock wave on J and x_α to denote the location of any other small wave α.

$$\begin{aligned} F(J) &\equiv |\sigma_J| + \sum_{j>1} \bar{L}_j(J) + M_1 \bar{L}_1(J) + M_2 Q(J), \\ \bar{L}_j(J) &\equiv \sum \{|\alpha| : \alpha \text{ is a small } j\text{-wave which crosses } J\}, \\ \sigma_J, |\beta_J| &\equiv \text{ the speed and strength of the strong shock crossing } J, \text{ respectively}, \\ Q(J) &\equiv Q_0(J) + Q_1(J) + Q_2(J), \end{aligned}$$

$$Q_0(J) \equiv \sum \{|\alpha\beta| : \alpha \text{ and } \beta \text{ are strengths of small waves which are approaching and cross } J\},$$

$$Q_1(J) \equiv \sum \left\{ \frac{|\alpha|}{\lambda_*^2} \int_{x_\alpha}^{\infty} G(x)dx : \alpha \text{ is the strength of 1-wave which crosses } J \text{ with } x_\alpha < x_J \right\},$$

$$+ \sum \left\{ \frac{|\alpha|}{\lambda_*^2} \int_{x_J}^{x_\alpha} G(x)dx + \frac{|\alpha|}{\lambda_*^2} \int_{x_J}^{\infty} G(x)dx : \alpha \text{ is the strength of 1-wave which crosses } J \text{ with } x_\alpha > x_J \right\},$$

$$Q_2(J) \equiv \sum_{i>1} \sum \left\{ \frac{|\alpha|}{\lambda_*^2} \int_{x_\alpha}^{\infty} G(x)dx : \alpha \text{ is the strength of } i\text{-wave which crosses for } J \text{ for } i \neq 1 \right\},$$

where $1 \ll M_1 \ll M_2$ are positive constants which will be determined later. Here λ_* is the lower bound on $|\lambda_1(u)|$ defined before Theorem 1.2.

Remark 2.1. It is noted in [14] that the linear part of the Glimm-type functional is $L(J) = |\sigma_J| + \bar{L}(J)$. However, here we use $L(J) = |\sigma_J| + \sum_{j\neq 1} \bar{L}_j(J) + M_1\bar{L}_1(J)$. The reason for this is as follows. Near the relatively strong shock, small 1-waves from the perturbation will be combined with the relatively strong shock so that $L_1(J)$ decreases by the amount of the total strengths of these small 1-waves up to the order of the interaction potential. On the other hand, the interaction of the small 1-wave α with the relatively strong shock changes the speed of the relatively strong shock by the amount of $|\alpha|$. Therefore, by choosing $M_1 \ll M_2$ large enough to compensate for the contribution to the change of the speed of the relatively strong shock and the new created waves, $F(J)$ can be shown to be decreasing in time by the usual argument (cf. [14]).

Next, for the convenience of the readers, we will briefly explain the simplified wave pattern given in [11, 21, 22]. Let $\epsilon > 0$ be small, and let time $T > 0$ be given. Set $N = \frac{1}{\epsilon}$, and choose M such that

$$(M-1)Ns < T \leq MNs.$$

Without loss of generality, we may assume that N and M are integers; then it is easy to see that, for a fixed N,

$$\lim_{s\to 0} M = \infty.$$

Let δ be the error from the randomness of an equidistributed sequence $\{a_j\}$. Then we have

$$\delta \to 0, \quad \text{as} \quad M \to \infty, \quad \text{for any } \epsilon.$$

In the following, we define a time zone Λ_p, an interaction measure $Q(\Lambda_p)$, and a cancellation measure $C(\Lambda_p)$. For $i = 1, \ldots, M$,

$$\Lambda_p \equiv \{(x,t) \in \mathbb{R} \times \mathbb{R}_+ : (p-1)Ns \le t < pNs\},$$
$$Q(\Lambda_p) = \sum_{\Delta_{mn} \in \Lambda_p} Q(\Delta_{mn}), \quad C(\Lambda_p) = \sum_{\Delta_{mn} \in \Lambda_p} C(\Delta_{mn}),$$

where a local interaction measure $Q(\Lambda_{mn})$ and a local cancellation measure $C(\Lambda_{mn})$ are defined as follows. Let Δ_{mn} be a diamond whose vertices are $((m-1)r + a_j r, ns), (mr + a_j r, ns), (mr, (n+\frac{1}{2})s)$, and $(mr, (n-\frac{1}{2})s)$.

$$Q(\Delta_{mn}) \equiv \sum_{(\alpha_i,\beta_j):opp} \{|\alpha_i||\beta_j| : \ \alpha_i \text{ and } \beta_j \text{ pass through } \Delta_{mn} \ \},$$
$$C(\Delta_{mn}) \equiv \sum_{(\alpha_i,\beta_j):opp} \left\{ \frac{|\alpha_i| + |\beta_i| - |\alpha_i + \beta_j|}{2} : \ \alpha_i \text{ and } \beta_i \text{ pass through } \Delta_{mn} \right\},$$

where $(\alpha_i, \beta_j) : opp$ denotes the approaching pair defined in section 3.1 in [11].

In Λ_p, we can partition all waves at time $t = (p-1)Ns$ into surviving waves and cancelled waves as in [11, 21, 22]. Based on this wave partition, we can define a simplified wave pattern which consists only of surviving nonlinear waves with fixed speeds in each small time zone Λ_p. That is, in Λ_p all nonlinear waves are linearly superimposed. More precisely, since $u_r(x,t)$ is of bounded variation, for a given small number ϵ, we can find E such that $T.V\{u_r(x,t) : |x| > E\} < \epsilon$ for $(p-1)Ns \le t \le pNs$. Then we replace the $u_r(x,t)$ on $x < -E$ or $x > E$ by the values of $\lim_{x\to-\infty} u(x,t)$ or $\lim_{x\to\infty} u(x,t)$, respectively, in Λ_p. Therefore, we have a finite number of surviving waves in Λ_p. Let us denote surviving i-waves in Λ_p by $v_i^1, \dots, v_i^N$. For each i-wave v_i^k, its location in the Glimm scheme is randomly chosen by the sequence $\{a_j\}$. However, in the simplified wave pattern, it is replaced by the line connecting its locations at time $t = (p-1)Ns$ and $t = pNs$. As in the approximate solutions of the Glimm scheme, the hyperbolic waves in the simplified wave pattern are connected by stationary waves. Notice also that i-waves do not cross each other in Λ_p.

For the secondary waves such as nonsurviving waves in $u_r(x,t)$ and generated waves from the nonlinear interactions in Λ_p, we do not keep track of them in Λ_p but put them back at $t = pNs+$. This generates an error in the L^1-norm which vanishes eventually as $s \to 0$. Therefore, we can assume that waves in the simplified wave pattern $\bar{u}_r(x,t)$ move in a deterministic way, but their end states evolve according to a stationary solution. For details, please refer to [11, 21, 22].

Let $u(x,t)$ and $v(x,t)$ be two Glimm solutions of (1.1) such that

$$\lim_{r\to 0} u_r(x,t) = u(x,t), \quad \lim_{r\to 0} v_r(x,t) = v(x,t) \quad \text{in } L^1_{loc}(\mathbb{R} \times \mathbb{R}_+),$$

and let $\bar{u}_r(x,t)$ and $\bar{v}_r(x,t)$ be the corresponding simplified wave patterns, respectively. For the time being, we will fix r, and, without any confusion, we denote $\bar{u}_r(x,t)$ and $\bar{v}_r(x,t)$ by $u(x,t)$ and $v(x,t)$, respectively. For given $(x,t) \in \mathbb{R} \times \mathbb{R}_+$, we resolve a discontinuity $(u(x,t), v(x,t))$ by the n Hugoniot curves; i.e.,

$$\omega_0(x,t) = u(x,t), \quad \omega_n(x,t) = v(x,t), \quad \omega_i(x,t) \in H_i(\omega_{i-1}(x,t)), \ i = 0, 1, \dots, n.$$

Here, $H_i(\omega_{i-1}(x,t))$ is an ith Hugoniot curve passing through $\omega_{i-1}(x,t)$ in the state space. We define the strength $q_i(x,t)$ of $(\omega_{i-1}(x,t), \omega_i(x,t))$ as follows:

$$q_i(x,t) \equiv l_i(u_0) \cdot (\omega_i(x,t) - \omega_{i-1}(x,t)), \quad i = 1, \dots, n.$$

The strict hyperbolicity implies

$$\frac{1}{C_1}|u(x,t)-v(x,t)| \le \sum_{i=1}^{n}|q_i(x,t)| \le C_1|u(x,t)-v(x,t)| \tag{2.2}$$

for some positive constant C_1 independent of t.

In what follows, we will use the following simplified notation:

$$\begin{aligned}\mathcal{J}(u),\ \mathcal{J}(v):&\ \text{the set of all discontinuities in } u(x,t), v(x,t),\ \text{respectively,}\\ &\ \text{and } \mathcal{J}=\mathcal{J}(u)\cup\mathcal{J}(v).\end{aligned}$$

For an i-wave $\alpha^i \in \mathcal{J}$, we define the location of α^i and the virtual waves $q_j^{\pm}(\alpha^i)$ generated by the difference of u and v at both sides of α^i as follows:

$$\begin{aligned}x(\alpha^i) &\equiv \ \text{the location of an } i\text{-wave } \alpha^i,\\ q_j^{\pm}(\alpha^i) &\equiv q_j(x(\alpha^i)\pm,t),\quad \lambda_j^{\pm}(\alpha^i)\equiv\lambda_j(\omega_{j-1}(x(\alpha^i)\pm,t),\omega_j(x(\alpha^i)\pm,t)).\end{aligned}$$

In the following, we state three lemmas which are direct consequences of the smoothness of the Hugoniot curves which are proved in [5] and [22].

LEMMA 2.1. *Let $\bar{u}\in\mathcal{N}$ and $k\in\{1,2,\dots,n\}$. Let us define the states and wave speeds as follows:*

$$\begin{aligned}&u=H_k(\xi)(\bar{u}),\quad u'=H_k(\xi')(u),\quad u''=H_k(\xi+\xi')(\bar{u}),\\ &\lambda=\lambda_k(\bar{u},u),\quad \lambda'=\lambda_k(u,u'),\quad \lambda''=\lambda_k(\bar{u},u'').\end{aligned}$$

Then we have

$$|(\xi+\xi')\lambda''-(\xi\lambda+\xi'\lambda')| = \mathcal{O}(1)|\xi||\xi'||\xi+\xi'|.$$

LEMMA 2.2. *Suppose that ξ_j,ξ_j', and ξ_j'' satisfy*

$$H_n(\xi_n)\circ\cdots\circ H_1(\xi_1)(u)=H_n(\xi_n')\circ\cdots\circ H_1(\xi_1')\circ H_n(\xi_n'')\circ\cdots\circ H_1(\xi_1'')(u).$$

Then

$$\sum_{i=1}^{n}|\xi_i-\xi_i'-\xi_i''| = \mathcal{O}(1)\left\{\sum_{j=1}^{n}|\xi_j'||\xi_j''||\xi_j'+\xi_j''|+\sum_{j>i}|\xi_j''||\xi_i'|\right\}.$$

If the values ξ_i' and ξ are related by

$$R_i(\xi)(u^*)=H_n(\xi_n')\circ\cdots\circ H_1(\xi_1')(u^*),$$

then

$$|\xi-\xi_i'|+\sum_{j\neq i}|\xi_j'| = \mathcal{O}(1)\left\{|\xi||\xi_i'||\xi+\xi_i'|+\sum_{j\neq i}|\xi_j'||\xi|\right\}.$$

Suppose $\alpha^i=(v_-,v_+)\in\mathcal{J}$ is an i-wave in either v or u and the other solution is continuous at $x=x(\alpha^i)$.

Denote

$$\begin{aligned} e(\Lambda_p) &\equiv (Q(\Lambda_p) + C(\Lambda_p) + \delta + \epsilon + NsG_0), \\ \Gamma_s(\alpha^i) &\equiv |\alpha^i||q_i^-(\alpha^i)||q_i^+(\alpha^i)|, \\ \Gamma_d(\alpha^i) &\equiv |\alpha^i| \sum_{j>i} |q_j^-(\alpha^i)| + |\alpha^i| \sum_{j<i} |q_j^+(\alpha^i)|. \end{aligned}$$

The following lemma gives the variation of $q_j(x,t)$ across the wave α^i.

LEMMA 2.3. *Let $\alpha^i = (v_-, v_+) \in \mathcal{J}$ be an i-wave in the time zone Λ_p. Then we have*

$$q_j^+(\alpha^i) = \begin{cases} q_i^-(\alpha^i) + [\alpha^i] + \mathcal{O}(1)(\Gamma_s + \Gamma_d)(\alpha^i) + \mathcal{O}(1)|\alpha^i|e(\Lambda_p), & j = i, \\ q_j^-(\alpha^i) + \mathcal{O}(1)(\Gamma_s + \Gamma_d)(\alpha^i) + \mathcal{O}(1)|\alpha^i|e(\Lambda_p), & j \neq i, \end{cases}$$

where $[\alpha^i] \equiv l_i(u_0) \cdot (v_+ - v_-)$.

Finally, we define a nonlinear functional $H(t)$, which is the weighted sum of four component functionals: First, we define a linear part $L(t)$ by

$$L(t) = \sum_{j=1}^{n} L^j(t), \;\; L^j(t) \equiv \int_{-\infty}^{\infty} |q_j(x,t)| dx.$$

Then it follows from (2.2) that $L(t)$ is equivalent to $||u(\cdot,t) - v(\cdot,t)||_{L^1(\mathbb{R})}$. As shown in [5], [22], $L(t)$ may increase in time t. So, in order to compensate for the possible increase of $L(t)$, we need to consider functionals with a good decay property in time t. They are $Q_d(t)$ measuring nonlinear couplings between waves of different characteristic families, $E(t)$ capturing the nonlinearity of the characteristic field due to the bifurcation of a shock curve and a rarefaction curve, and $Q_{so}(t)$ measuring the source effect on the L^1 distance. Both $Q_d(t)$ and $E(t)$ are used in the study of homogeneous hyperbolic conservation laws (cf. [5, 20, 21, 22]). We now explain $Q_{so}(t)$ briefly. This functional is defined to capture the effect of the source on the L^1 distance. For this, we need to consider the potential interactions between the virtual waves $q_i(x,t)$ and stationary waves. For definiteness, we consider a virtual wave $q_j(x_0,t)$, $j \geq 2$, located at $x = x_0$. Since this wave has a positive speed, it will interact with the stationary waves lying $x \geq x_0$ with the interaction potential of the order

$$\sum_{j \geq 2} |q_j(x_0,t)| \left(\int_{x(q_j)}^{\infty} G(\xi) d\xi \right).$$

For the resonant field $(r_1(u), \lambda_1(u))$, the left states and right states of a relatively strong shock are supersonic and subsonic, respectively. Next we consider the potential interactions with an imaginary wave $q_1(x_0,t)$ and stationary waves. If $x_0 < x_f(t)$ (= location of the relatively strong shock at time t), then the imaginary wave $q_1(x_0,t)$ has a positive speed, so before it interacts with a relatively strong shock, it will interact with stationary waves, and after it interacts with a relatively strong shock, as part of a relatively strong wave, it will interact with the stationary wave lying on $[0,1]$, so the potential interactions with $q_1(x_0,t)$ and stationary waves are

$$|q_1(x_0,t)| \int_{x(q_1(x_0,t))}^{\infty} G(\xi) d\xi.$$

In contrast, if $x_0 > x_f(t)$, then the imaginary wave $q_1(x_0,t)$ has a negative speed and will interact with stationary waves before it is combined with a relatively strong shock wave, and after it is absorbed into a relatively strong shock wave, as part of a relatively strong shock, it will interact with stationary waves. So potential interactions are

$$|q_1(x_0,t)|\left(\int_{x(q_1(x_0,t))}^{x_f(t)} G(\xi)d\xi + \int_{x_f(t)}^{\infty} G(\xi)d\xi\right).$$

Based on this observation, we define a nonlinear functional $H(t)$ for two simplified wave patterns as follows:

$$\begin{aligned}
&L(t) = \sum_{j=1}^{n} L^j(t), \quad L^j(t) \equiv \int_{-\infty}^{\infty} |q_j(x,t)|dx,\\
&Q_d(t) = \sum_{\alpha^i \in \mathcal{J}} Q_d(\alpha^i(t)),\\
&Q_d(\alpha^i(t)) \equiv |\alpha^i(t)|\left\{\sum_{j>i}\int_{-\infty}^{x(\alpha^i)} |q_j(x,t)|dx + \sum_{j<i}\int_{x(\alpha^i)}^{\infty} |q_j(x,t)|dx\right\},\\
&E(t) = \sum_{\alpha^i \in \mathcal{J}} E(\alpha^i(t)),\\
&E(\alpha^i(t)) \equiv |\alpha^i(t)| \cdot \begin{cases} \int_{-\infty}^{x(\alpha^i)} q_i(x,t)_+dx + \int_{x(\alpha^i)}^{\infty} q_i(x,t)_-dx, \ \alpha^i \in J(u),\\ \int_{x(\alpha^i)}^{\infty} q_i(x,t)_+dx + \int_{-\infty}^{x(\alpha^i)} q_i(x,t)_-dx, \ \alpha^i \in J(v), \end{cases}\\
&Q_{so}(t) = \sum_{j\geq 1} Q_{so}^j(t),\\
&Q_{so}^j(t) \equiv \int_{-\infty}^{\infty} |q_j(x,t)|\left(\int_{x(q_j)}^{\infty} G(\xi)d\xi\right)dx, \ j \geq 2,\\
&Q_{so}^1(t) \equiv \int_{-\infty}^{x_f(t)} |q_1(x,t)|\left(\int_{x(q_1)}^{\infty} G(\xi)d\xi\right)dx\\
&\qquad + \int_{x_f(t)}^{\infty} |q_1(x,t)|\left(\int_{x_f(t)}^{x(q_1)} G(\xi)d\xi + \int_{x_f(t)}^{\infty} G(\xi)d\xi\right)dx,\\
&H(t) \equiv [1 + K_1F((p-1)Ns)]L(t) + K_2[Q_d(t) + E(t) + Q_{so}(t)], \ t \in [(p-1)Ns, pNs),\\
&\quad 1 \leq p \leq M,
\end{aligned}$$

where K_1 and K_2 are positive constants to be determined later and $F(t) = F(u(x,t)) + F(v(x,t))$ is sum of the Glimm functionals for the solutions $u(x,t)$ and $v(x,t)$. In the following sections, we will study the time-evolutional property of this nonlinear functional $H(t)$ under the condition (1.3).

3. Basic estimates. In this section, we study some basic estimates which are necessary for the decay analysis of the nonlinear functional $H(t)$. First, we estimate the time variation of a shock strength through a stationary background.

LEMMA 3.1. *Let $\alpha^i(t) = (u_-(t), u_+(t))$, $t \in [(p-1)Ns, pNs)$ be an i-wave issued from $(hr, (p-1)Ns)$ in the simplified wave pattern $u(x,t)$. Then*

$$\frac{d|\alpha^i(t)|}{dt} = \mathcal{O}(1)G(x(\alpha^i))|\alpha^i(t)|,$$

where $\mathcal{O}(1)$ depends only on (1.1).

Proof. Let $(x(t), t)$ be the locus of the shock in x-t plane. Denote the states at both sides of the shock as follows:

$$u_-(t) \equiv u(x(t)-, t), \quad u_+(t) \equiv u(x(t)+, t).$$

Since $u(x,t)$ is a local steady state solution, we have

$$u_x = (f'(u))^{-1} g(x,u) \equiv \tilde{g}(x,u). \tag{3.1}$$

From this, we have

$$u_+(t+h) = u_+(t) + \int_{x(t)}^{x(t+h)} \tilde{g}(\xi, u_+(\xi))d\xi, \tag{3.2}$$

$$u_-(t+h) = u_-(t) + \int_{x(t)}^{x(t+h)} \tilde{g}(\xi, u_-(\xi))d\xi. \tag{3.3}$$

It follows from (3.2) and (3.3) that

$$u_+(t+h) - u_-(t+h) = u_+(t) - u_-(t) + \mathcal{O}(1) \int_{x(t)}^{x(t+h)} G(\xi)|\alpha^i(t)|d\xi,$$

where we have used the fact that $|u_+(\xi) - u_-(\xi)| = \mathcal{O}(1)|\alpha^i(t)|$. Hence we have

$$(u_-(t+h), u_+(t+h))_i = (u_-(t), u_+(t))_i + \mathcal{O}(1) \int_{x(t)}^{x(t+h)} G(\xi)|\alpha^i(t)|d\xi.$$

This implies that

$$\frac{d|\alpha^i(t)|}{dt} = \mathcal{O}(1) G(x(\alpha^i))|\alpha^i(t)|,$$

which completes the proof. ∎

Assume now at time t that there is no wave interaction. Let $\mathcal{J} = \{\alpha_j\}_{j=1}^{N_t}$ be the set of all waves in u and v at this time with locations

$$-\infty < x(\alpha_1) < \cdots < 0 \le x(\alpha_k) < x(\alpha_{k+1}) < \cdots < 1 \le x(\alpha_l) < \cdots < x(\alpha_{N_t}) < \infty.$$

Without loss of generality, we may assume that $x(\alpha_k(t)) = 0$ and $x(\alpha_l(t)) = 1$. As shown in [11], the term

$$II^j \equiv \sum_{\alpha \in J} \{\lambda(q_j^-(\alpha))|q_j^-(\alpha)| - \lambda(q_j^+(\alpha))|q_j^+(\alpha)|\}$$

denotes the effect of the source on the time evolution of $L^j(t)$. In the following lemma, we estimate this quantity.

LEMMA 3.2. *For each $j, k \in \{1, \ldots, n\}$, we have the following estimates.*

$$f^j(\omega_k(x,t))_x - g^j(x, \omega_k(x,t)) = \mathcal{O}(1) G(x) \max\left(\sum_{i=1}^{k} |q_i(x,t)|, \sum_{i=k+1}^{n} |q_i(x,t)| \right).$$

Proof. For given smooth functions $p(x,t)$ and $q(x,t)$, we set

$$h^j(x,p,q) \equiv \nabla_p f^j(p)q - g^j(x,p).$$

Then, since $u(x,t) = \omega_0(x,t)$ and $v(x,t) = \omega_n$ are local steady state solutions,

$$h^j(x,\omega_0,\partial_x\omega_0) = h^j(x,\omega_n,\partial_x\omega_n) = 0. \tag{3.4}$$

By the mean value theorem and (3.4), we have

$$\begin{aligned} h^j(x,\omega_k,\partial_x\omega_k) &= h^j(x,\omega_k,\partial_x\omega_k) - h^j(x,\omega_0,\partial_x\omega_0) \\ &= h^j(x,\omega_k,\partial_x\omega_k) - h^j(x,\omega_0,\partial_x\omega_k) + h^j(x,\omega_0,\partial_x\omega_k) - h^j(x,\omega_0,\partial_x\omega_0) \\ &= \nabla_p h^j(x,\theta_k^1,\partial_x\omega_k)\cdot(\omega_k-\omega_0) + \nabla_q h^j(x,\omega_0,\theta_k^2)\partial_x(\omega_k-\omega_0), \end{aligned} \tag{3.5}$$

where θ_k^1 is a point on the line segment connecting ω_0 and ω_k and θ_k^2 is a point on the line segment connecting $\partial_x\omega_0$ and $\partial_x\omega_k$, respectively. Next, we claim the following:

1. $\nabla_p h^j(x,\theta_k^1,\partial_x\omega_k) = \mathcal{O}(1)G(x)$.
2. $\nabla_q h^j(x,\omega_0,\theta_k^2) = \mathcal{O}(1)$, $\partial_x(\omega_k-\omega_0) = \mathcal{O}(1)G(x)\max(\sum_{i=1}^k |q_i(x,t)|, \sum_{i=k+1}^n |q_i(x,t)|)$.

Proof of 1. Since $\omega_k(x,t)$ is determined by connecting two end states $u(x,t)$ and $v(x,t)$ by Hugoniot curves, we can write

$$\omega_k(x,t) = \Phi^k(u(x,t),v(x,t)) \quad \text{for some } C^2 \text{ function } \Phi^k.$$

It follows from (3.1) that

$$\partial_x(\omega_k(x,t)) = \Phi_u^k u_x + \Phi_v^k v_x = \mathcal{O}(1)G(x), \tag{3.6}$$

where we have used the fact that all states are not sonic, i.e., $\lambda_i(u) \neq 0$ for all i and $u \in \mathcal{N}$. On the other hand, (3.6) implies

$$\begin{aligned} \nabla_p h^j(x,\theta_k^1,\partial_x\omega_k) &= \nabla_p^2 f^j(\theta_k^1)\partial_x\omega_k - \nabla_p g^j(x,\theta_k^1) \\ &= \mathcal{O}(1)G(x). \end{aligned}$$

Proof of 2. By definition of h^j, it is easy to see

$$\nabla_q h^j(x,\omega_0,\theta_k^2) = \mathcal{O}(1).$$

Next we show

$$\partial_x(\omega_k-\omega_0) = \mathcal{O}(1)G(x)\max\left(\sum_{i=1}^k |q_i(x,t)|, \sum_{i=k+1}^n |q_i(x,t)|\right).$$

For given t, we define $\omega_k^*(y,t)$ by a local steady state solution such that

$$f(\omega_k^*(x,t))_x = g(x,\omega_k^*(x,t)), \quad \omega_k^*(x,t) = \omega_k(x,t).$$

Notice that when $\sum_{i=k+1}^n |q_i(x,t)| = 0$, since $v(x,t) = \omega_k(x,t)$, by the local uniqueness of the ODE solution we have

$$\omega_k^*(x,t) = v(x,t).$$

In this case, the difference between $\omega_k(x+\Delta x,t)$ and $\omega_k^*(x+\Delta x,t)=v(x+\Delta x,t)$ is due to the interaction between the imaginary waves $q_i(x,t), i=1,\dots,k$, and stationary waves. Therefore, the difference is the order of the interaction errors between imaginary waves $q_i(x,t),\ i=1,\dots,k$, and stationary waves, i.e.,

$$\omega_k(x+\Delta x,t)-\omega_k^*(x+\Delta x,t)=\mathcal{O}(1)\max\sum_{i=1}^{k}|q_i(x,t)|\int_x^{x+\Delta x}G(\eta)d\eta. \tag{3.7}$$

By the same argument as above, when $\sum_{i=1}^{k}|q_i(x,t)|=0$, we have

$$\omega_k(x+\Delta x,t)-\omega_k^*(x+\Delta x,t)=\mathcal{O}(1)\max\sum_{i=1}^{k}|q_i(x,t)|\int_x^{x+\Delta x}G(\eta)d\eta. \tag{3.8}$$

By (3.7), (3.8), and the continuity argument, we obtain

$$\begin{aligned}&\omega_k(x+\Delta x,t)-\omega_k^*(x+\Delta x,t)\\ &\quad=\mathcal{O}(1)\max\left(\sum_{i=1}^{k}|q_i(x,t)|,\sum_{i=k+1}^{n}|q_i(x,t)|\right)\int_x^{x+\Delta x}G(\eta)d\eta.\end{aligned} \tag{3.9}$$

On the other hand, we can rewrite

$$\partial_x[\omega_k(x,t)-\omega_0(x,t)]=\partial_x[\omega_k(x,t)-\omega_k^*(x,t)]+\partial_x[\omega_k^*(x,t)-\omega_0(x,t)]=\Gamma_1+\Gamma_2.$$

Then, by the same argument as in [14], we have

$$\Gamma_2=\mathcal{O}(1)\sum_{i=1}^{k}|q_i(x,t)|G(x).$$

Now, we estimate Γ_1 using (3.9). Since $\omega_k(x,t)-\omega_k^*(x,t)=0$, we have

$$\begin{aligned}&[\omega_k(x+\Delta x,t)-\omega_k^*(x+\Delta x,t)]-[\omega_k(x,t)-\omega_k^*(x,t)]\\ &=\mathcal{O}(1)\max\left(\sum_{i=1}^{k}|q_i(x,t)|,\sum_{i=k+1}^{n}|q_k(x,t)|\right)\int_x^{x+\Delta x}G(\eta)d\eta.\end{aligned}$$

Now, dividing by Δx and letting $\Delta x\to 0$, we obtain

$$\partial_x[\omega_k(x,t)-\omega_k^*(x,t)]=\mathcal{O}(1)\max\left(\sum_{i=1}^{k}|q_i(x,t)|,\sum_{i=k+1}^{n}|q_i(x,t)|\right)G(x).$$

Combining the estimates for Γ_1 and Γ_2, we have

$$\partial_x(\omega_k(x,t)-\omega_0(x,t))=\mathcal{O}(1)\max\left(\sum_{i=1}^{k}|q_i(x,t)|,\sum_{i=k+1}^{n}|q_i(x,t)|\right)G(x).$$

In (3.5), using the above claim, we obtain

$$f^j(\omega_k(x,t))_x-g^j(x,\omega_k(x,t))=\mathcal{O}(1)G(x)\max\left(\sum_{i=1}^{k}|q_i(x,t)|,\sum_{i=k+1}^{n}|q_i(x,t)|\right).$$

This completes the proof. □

LEMMA 3.3. *Suppose that* $G_{11}(x,u) \le -\lambda$ $(\lambda > 0)$. *Then we have*

$$\sum_{j=1}^{n} II^j \le \mathcal{O}(1)\sum_{k=1}^{n}\int_{-\infty}^{\infty} G(x)|q_k(x,t)|dx - |\mathcal{O}(1)|\lambda\int_{-\infty}^{\infty}\mathbf{1}_{G(x)>0}|q_1(x,t)|dx.$$

Proof. Let us set

$$II^j \equiv \sum_{i=1}^{N_t-1} II^j(\alpha_i,\alpha_{i+1}), \quad II^j(\alpha_i,\alpha_{i+1}) \equiv \lambda(q_j^-(\alpha_{i+1}))|q_j^-(\alpha_{i+1})| - \lambda(q_j^+(\alpha_i))|q_j^+(\alpha_i)|.$$

We first estimate $II^j(\alpha_i,\alpha_{i+1})$. Since $\lambda(q_j(x))|q_j(x)|$ is continuous on $(x(\alpha_i),x(\alpha_{i+1}))$, we have

$$\lambda(q_j^+(x,t))|q_j^+(x,t)| = \lambda(q_j^-(x,t))|q_j^-(x,t)|, \quad x \in (x(\alpha_i),x(\alpha_{i+1})).$$

So, if necessary, by inserting

$$\lambda(q_j^+(x,t))|q_j^+(x,t)| - \lambda(q_j^-(x,t))|q_j^-(x,t)|, x \in (x(\alpha_i),x(\alpha_{i+1})),$$

it suffices to consider only two cases, i.e., either $q_j(x,t) \ge 0$ or $q_j(x,t) < 0$, on the whole interval $(x(\alpha_i),x(\alpha_{i+1}))$.

In the following, we consider only the case in which $q_j(x,t) \ge 0$ on $(x(\alpha_i),x(\alpha_{i+1}))$. The other case can be discussed similarly. By a direct calculation, we have

$$\begin{aligned} II^j(\alpha_i,\alpha_{i+1}) &= f^j(\omega_j(x(\alpha_{i+1}))) - f^j(\omega_j(x(\alpha_i))) \\ &\quad - [f^j(\omega_{j-1}(x(\alpha_{i+1}))) - f^j(\omega_{j-1}(x(\alpha_i)))] \\ &= \int_{x(\alpha_i)}^{x(\alpha_{i+1})} \left[f^j(\omega_j(x,t))_x - f^j(\omega_{j-1}(x,t))_x\right] dx. \end{aligned} \tag{3.10}$$

And now we can prove the estimates in the lemma as follows.

Case 1. $j = 1$. By Lemma 3.2, we have

$$\begin{aligned} &f^1(\omega_1(x,t))_x = g^1(x,\omega_1(x,t)) + \mathcal{O}(1)G(x)\max\left(|q_1(x,t)|, \sum_{k=2}^{n}|q_k(x,t)|\right), \\ &f^1(\omega_0(x,t))_x = g^1(x,\omega_0(x,t)). \end{aligned}$$

From (3.10), we have

$$\begin{aligned} II^1(\alpha_i,\alpha_{i+1}) &= \int_{x(\alpha_i)}^{x(\alpha_{i+1})} \left[g^1(x,\omega_1(x,t)) - g^1(x,\omega_0(x,t))\right] dx \\ &\quad + \mathcal{O}(1)\int_{x(\alpha_i)}^{x(\alpha_{i+1})} G(x)\left(\sum_{k=1}^{n}|q_k(x,t)|\right) dx. \end{aligned} \tag{3.11}$$

Moreover, we have

$$\begin{aligned} g^1(x,\omega_1(x,t)) - g^1(x,\omega_0(x,t)) &= \nabla_u g^1(x,\bar{\theta}_1(x,t))\cdot(\omega_1(x,t) - \omega_0(x,t)) \\ &= \nabla_u g^1(x,\bar{\theta}_1(x,t))\cdot\sum_{i=1}^{n}(l_i(u_0)\cdot(\omega_1(x,t) - \omega_0(x,t)))r_i(u_0) \end{aligned}$$

$$= \left(l_1(u_0) \frac{\partial g(x, \bar{\theta}_1)}{\partial u} r_1(u_0) \right) |q_1(x,t)| + \nabla_u g^1(x, \bar{\theta}_1(x,t)) \cdot \sum_{i=2}^{n} [l_i(u_0) \cdot (\omega_1(x,t) - \omega_0(x,t))] r_i(u_0),$$

where $\bar{\theta}_1(x,t)$ is a point on the line segment connecting $\omega_0(x,t)$ and $\omega_1(x,t)$. On the other hand, we assume that d ($\equiv$ the diameter of $\mathcal{N}$) is small. Since

$$\omega_1 = \omega_0 + q_1(x,t) r_1(u_0) + \mathcal{O}(1) d |q_1(x,t)|,$$

we therefore have, for $i \neq 1$,

$$\begin{aligned} l_i(u_0) \cdot ((\omega_1(x,t) - \omega_0(x,t))) &= q_1(x,t)[l_i(u_0) \cdot r_1(u_0)] + \mathcal{O}(1) d |q_1(x,t)| \\ &= \mathcal{O}(1) d |q_1(x,t)| \ll |q_1(x,t)|, \end{aligned}$$

where we have used that $l_i(u_0) \cdot r_1(u_0) = 0,\ i \neq 1$. Hence we have

$$g^1(x, \omega_1(x,t)) - g^1(x, \omega_0(x,t)) \leq |\mathcal{O}(1)| \left(l_1(u_0) \frac{\partial g(x, \bar{\theta}_1)}{\partial u} r_1(u_0) \right) |q_1(x,t)|. \tag{3.12}$$

Since $G_{11}(x,u) \leq -\lambda$ for some $\lambda > 0$, it follows from (3.11) and (3.12) that

$$\begin{aligned} II^1(\alpha_i, \alpha_{i+1}) \leq\ & \mathcal{O}(1) \int_{x(\alpha_i)}^{x(\alpha_{i+1})} G(x) \left(\sum_{k=1}^{n} |q_k(x,t)| \right) dx \\ & - |\mathcal{O}(1)| \lambda \int_{x(\alpha_i)}^{x(\alpha_{i+1})} \mathbf{1}_{G(x)>0} |q_1(x,t)| dx, \end{aligned}$$

where $\mathbf{1}_{G(x)>0}$ denotes the characteristic function of the set $\{G(x) > 0\}$. By summing up all $II^1(\alpha_i, \alpha_{i+1})$ over all $i = 1, \ldots, N_t - 1$, we have

$$II^1 \leq \mathcal{O}(1) \int_{-\infty}^{\infty} G(x) \left(\sum_{k=1}^{n} |q_k(x,t)| \right) dx - |\mathcal{O}(1)| \lambda \int_{-\infty}^{\infty} \mathbf{1}_{G(x)>0} |q_1(x,t)| dx.$$

Case 2. By Lemma 3.2, we have

$$\sum_{j=2}^{n} II^j = \mathcal{O}(1) \int_{-\infty}^{\infty} G(x) \left(\sum_{k=1}^{n} |q_k(x,t)| \right) dx.$$

Combining Cases 1 and 2, we have

$$\sum_{j=1}^{n} II^j \leq \mathcal{O}(1) \int_{-\infty}^{\infty} G(x) \left(\sum_{k=1}^{n} |q_k(x,t)| \right) dx - |\mathcal{O}(1)| \lambda \int_{-\infty}^{\infty} \mathbf{1}_{G(x)>0} |q_1(x,t)| dx.$$

This completes the proof. ∎

Remark 3.1. In the resonant scalar case, the dissipation condition $G_{11}(x,u) \leq -\lambda$ gives an estimate

$$II^1 \leq -\lambda \int_{-\infty}^{\infty} \mathbf{1}_{G(x)>0} |q_1(x,t)| dx.$$

4. Stability analysis. In this section, we study the L^1 stability of a steady transonic shock wave solution under the condition (1.3). First, we study time decay rates for each component functional. Recall that an open interval $I_p = ((p-1)Ns, pNs), p \in \{1, \dots, M\}$, is the union of two disjoint sets, $I_p = I_p^1 \cup I_p^2$, where I_p^1 is the set of all countable interaction times such that $H(t)$ is simply continuous, and I_p^2 is the set of all differentiable points of $H(t)$. For notational convenience, we set

$$\Gamma_s \equiv \sum_{\alpha\in\mathcal{J}} \Gamma_s(\alpha), \quad \Gamma_d \equiv \sum_{\alpha\in\mathcal{J}} \Gamma_d(\alpha),$$
$$\Gamma_{so} \equiv \sum_{j=1}^{n} \int_0^1 G(x)|q_j(x,t)|dx, \quad \Gamma \equiv \Gamma_s + \Gamma_d + \Gamma_{so}.$$

Let $v_0(x)$ be a small perturbation of $u(x)$ such that

$$T.V_x(v_0(x) - u(x)) \ll 1.$$

In the following, we briefly sketch the time-evolution estimates of the Glimm functional defined in section 2. For the details, we refer to [14].

LEMMA 4.1 (see [14, 18]). *Assume that the condition* (1.3) *holds. Then we have the following estimates on the Glimm functional defined in section* 2*:*

$$F(J) \le F(0) - \frac{1}{2}Q(\Lambda_J) - \frac{c_0}{2}\sum_{k=1}^{k_J}\left|\int_{x_f((k-1)s)}^{x_f(ks)} G(x)dx\right|,$$

where c_0 is a positive constant.

Proof. Let J_1 and J_2 be space-like curves such that J_2 is an immediate successor of J_1 and $\Delta \equiv \Delta(J_1, J_2)$ is the diamond spanned by J_1 and J_2. Based on the estimates obtained in [14], we consider the following two cases.

Case 1. $\Delta(J_1, J_2)$ contains the relatively strong shock. In this case, we have

$$\begin{aligned}
|\sigma_{J_2}| - |\sigma_{J_1}| &\le -|\mathcal{O}(1)|\left|\int_{x_f(J_1)}^{x_f(J_2)} G(x)dx\right| + \mathcal{O}(1)\Big(Q(\Delta)\\
&\quad + \sum\{|\alpha_1| : \alpha_1 \text{ is a 1-wave passing through } J_1 \text{ in } \Delta\}\Big),\\
\bar{L}_1(J_2) - \bar{L}_1(J_1) &\le -\sum\{|\alpha_1| : \alpha_1 \text{ is a 1-wave passing through } J_1 \text{ in } \Delta\}\\
&\quad + \mathcal{O}(1)Q(\Delta),\\
\sum_{j\neq 1} L_j(J_2) - \sum_{j\neq 1} L_j(J_1) &\le \mathcal{O}(1)Q(\Delta),\\
Q(J_2) - Q(J_2) &\le -Q(\Delta) + \mathcal{O}(1)(\bar{L}(J_1) + G_1)\\
&\quad \cdot\left(Q(\Delta) + |\beta_{J_1}|\left|\int_{x_f(J_1)}^{x_f(J_2)} G(x)dx\right|\right).
\end{aligned}$$

By definition of $F(J)$, we have

$$\begin{aligned}
F(J_2) - F(J_1) &\le [-|\mathcal{O}(1)| + \mathcal{O}(1)M_2(\bar{L}(J_1) + G_1)|\beta_{J_1}|]\left|\int_{x_f(J_1)}^{x_f(J_2)} G(x)dx\right|\\
&\quad + (\mathcal{O}(1) + M_1\mathcal{O}(1) - M_2)Q(\Delta)
\end{aligned}$$

$$+ (\mathcal{O}(1) - M_1) \sum \{|\alpha_1| : \alpha_1 \text{ is a 1-wave } \in \Delta \cap J_1\}.$$

Since $1 \ll M_1 \ll M_2$ (see section 2), we have

$$-|\mathcal{O}(1)| + \mathcal{O}(1)M_2(\bar{L}(J_1) + G_1)|\beta_{J_1}| \le -c_0 \quad \text{for some positive constant } c_0,$$
$$\mathcal{O}(1) + M_1\mathcal{O}(1) - M_2 \le -\frac{1}{2}, \quad \mathcal{O}(1) - M_1 < 0.$$

Then we have

$$F(J_2) - F(J_1) \le -\frac{1}{2}Q(\Delta) - \frac{c_0}{2}\left|\int_{x_f(J_1)}^{x_f(J_2)} G(x)dx\right|.$$

Case 2. $\Delta(J_1, J_2)$ does not contain the relatively strong shock. We have

$$\begin{aligned} |\sigma_{J_2}| - |\sigma_{J_1}| &= 0, \\ \bar{L}_1(J_2) - \bar{L}_1(J_1) &\le \mathcal{O}(1)Q(\Delta), \\ \sum_{j\neq 1} L_j(J_2) - \sum_{j\neq 1} L_j(J_1) &\le \mathcal{O}(1)Q(\Delta), \\ Q(J_2) - Q(J_2) &\le -Q(\Delta) + \mathcal{O}(1)(\bar{L}(J_1) + G_1)Q(\Delta). \end{aligned}$$

Thus

$$F(J_2) - F(J_1) \le (\mathcal{O}(1) + M_1\mathcal{O}(1) - M_2)Q(\Delta).$$

Since $1 \ll M_1 \ll M_2$ in section 2, we have

$$\mathcal{O}(1) + M_1\mathcal{O}(1) - M_2 \le -\frac{1}{2},$$

which implies

$$F(J_2) - F(J_1) \le -\frac{1}{2}Q(\Delta).$$

By telescoping the above estimates from every space-like curve between J and 0, we obtain

$$F(J) \le F(0) - \frac{1}{2}Q(\Lambda_J) - \frac{c_0}{2}\sum_{k=1}^{k_J}\left|\int_{x_f((k-1)s)}^{x_f(ks)} G(x)dx\right|. \qquad \square$$

Remark 4.1. Because of the dissipation condition on $G_{11}(x, u)$, the speed of the relatively strong shock is decreasing in time when the relatively strong shock only interacts with stationary waves (see Lemma 3.1 in [14]), and this gives a good term,

$$-\sum_{k=1}^{k_J}\left|\int_{x_f((k-1)s)}^{x_f(ks)} G(x)dx\right|,$$

in Glimm functional $F(J)$. This good term will be used in Lemma 4.3 in order to control the bad term in $\frac{dQ^1_{so}(t)}{dt}$.

Next we estimate the time-evolution of $Q^1_{so}(t)$, which is different from $Q^1_{so}(t)$ in [11]. But the estimates for the $Q^i_{so}(t)$, $i \ge 2$, will be the same as in [11].

LEMMA 4.2. *Suppose that the main assumptions in section* 1 *hold. Then we have the following estimate:*

$$\frac{dQ^1_{so}(t)}{dt} \le -\lambda_* \int_{-\infty}^{\infty} G(x)|q_1(x,t)|dx + 2\left(\frac{d}{dt}\int_{x_f(t)}^{\infty} G(\xi)d\xi\right)\int_{x_f(t)}^{\infty}|q_1(x,t)|dx + \mathcal{O}(1)G_1\left\{\Gamma + (T.V + G_1)e(\Lambda_p)\right\},$$

where λ_* is a positive constant which is a lower bound on $|\lambda_1(u)|$.

Proof. For a given noninteraction time $t \in \Lambda_p$, we have three cases, depending on the location of a relatively strong shock wave:

$$x_f(t) < 0, \quad 0 \le x_f(t) \le 1, \quad x_f(t) > 1.$$

Case 1 ($x_f(t) < 0$). In this case, $Q^1_{so}(t)$ becomes

$$Q^1_{so}(t) = G_1\int_{-\infty}^{\infty}|q_1(x,t)|dx + \int_{x_f(t)}^{\infty}|q_1(x,t)|\left(\int_{x_f(t)}^{x(q_1)} G(\xi)d\xi\right)dx \equiv I_1 + I_2.$$

We denote the locations of waves as follows.

$$\begin{aligned}-\infty < x(\alpha_1) < \cdots < x(\alpha_e) = x_f(t) < x(\alpha_{e+1}) < \cdots < x(\alpha_k)\\ = 0 < x(\alpha_{k+1}) < \cdots < x(\alpha_l) = 1 < x(\alpha_{l+1}) < \cdots < x(\alpha_m) < \infty.\end{aligned}$$

For notational convenience, let us denote $x(\alpha_0) \equiv -\infty$ and $x(\alpha_{m+1}) \equiv \infty$.

Now, we take the derivative of $Q^1_{so}(t)$, and then we have

$$\frac{dQ^1_{so}(t)}{dt} = G_1\frac{d}{dt}\int_{-\infty}^{\infty}|q_1(x,t)|dx + \frac{d}{dt}\int_{x_f(t)}^{\infty}|q_1(x,t)|\left(\int_{x_f(t)}^{x(q_1)} G(\xi)d\xi\right)dx \equiv \frac{dI_1}{dt} + \frac{dI_2}{dt}. \tag{4.1}$$

By the same estimates in [14], the first term of (4.1) can be estimated as follows:

$$\frac{dI_1}{dt} = \mathcal{O}(1)G_1[\Gamma + (T.V + G_1)e(\Lambda_p)]. \tag{4.2}$$

Now, we consider the second term of (4.1). Using $\int_{x_f(t)}^{x} G(\xi)d\xi = 0,\ x \le 0$, we have

$$\frac{dI_2}{dt} = \sum_{i=k+1}^{l}\frac{d}{dt}\int_{x(\alpha_{i-1})}^{x(\alpha_i)}|q_1(x,t)|\left(\int_{x_f(t)}^{x(q_1)} G(\xi)d\xi\right)dx + G_1\sum_{i=l+1}^{m+1}\frac{d}{dt}\int_{x(\alpha_{i-1})}^{x(\alpha_i)}|q_1(x,t)|dx.$$

Then, by a direct calculation, we have

$$\begin{aligned}\frac{dI_2}{dt} &\le \sum_{i=k}^{l}\left(\int_{x_f(t)}^{x(\alpha_i))} G(\xi)d\xi\right)\dot{x}(\alpha_i)(|q_1^-(\alpha_i)| - |q_1^+(\alpha_i)|)\\ &\quad + G_1\sum_{i=l+1}^{m}\dot{x}(\alpha_i)(|q_1^-(\alpha_i)| - |q_1^+(\alpha_i)|) + \int_0^1|q_1(x,t)|\dot{x}(q_1)G(x)dx\\ &\le \mathcal{O}(1)G_1\left\{\sum_{i=k}^{m}(\Gamma_s + \Gamma_d)(\alpha_i) + \Gamma_{so} + (T.V + G_1)e(\Lambda_p)\right\}\\ &\quad - \lambda_*\int_0^1 G(x)|q_1(x,t)|dx,\end{aligned} \tag{4.3}$$

where we have used the fact that

$$G(x_f(t)) = 0, \quad \dot{x}(q_1) \le -\lambda_* \quad \text{if } x(q_1) > x_f(t).$$

Combining (4.2) and (4.3), we have

$$\frac{dQ^1_{so}(t)}{dt} \le \mathcal{O}(1)G_1(\Gamma + (T.V + G_1)e(\Lambda_p)) - \lambda_* \int_0^1 G(x)|q_1(x,t)|dx. \tag{4.4}$$

Case 2 ($0 \le x_f(t) \le 1$). We consider only the case for $0 < x_f(t) < 1$. The cases for $x_f(t) = 0$ and $x_f(t) = 1$ can be treated similarly. We denote the locations of waves as follows:

$$\begin{aligned} -\infty < x(\alpha_1) < \cdots < x(\alpha_k) = 0 < x(\alpha_{k+1}) < \cdots < x(\alpha_e) = x_f(t) \\ < x(\alpha_{e+1}) < \cdots < x(\alpha_l) = 1 < x(\alpha_{l+1}) < \cdots < x(\alpha_m) < \infty. \end{aligned}$$

For notational convenience, let us denote $x(\alpha_0) \equiv -\infty$ and $x(\alpha_{m+1}) \equiv \infty$.

Using $\int_x^\infty G(\xi)d\xi = G_1,\ x \le 0$, we have

$$\begin{aligned} \frac{dQ^1_{so}(t)}{dt} &= G_1 \sum_{i=1}^{k} \frac{d}{dt} \int_{x(\alpha_{i-1})}^{x(\alpha_i)} |q_1(x,t)|dx + \sum_{i=k+1}^{e} \frac{d}{dt} \int_{x(\alpha_{i-1})}^{x(\alpha_i)} |q_1(x,t)| \\ &\quad \cdot \left(\int_{x(q_1)}^{\infty} G(\xi)d\xi \right) dx \\ &\quad + \sum_{i=e+1}^{m+1} \frac{d}{dt} \int_{x(\alpha_{i-1})}^{x(\alpha_i)} |q_1(x,t)| \left(\int_{x_f(t)}^{x(q_1)} G(\xi)d\xi + \int_{x_f(t)}^{\infty} G(\xi)d\xi \right) dx. \end{aligned}$$

Then, by a direct calculation, we have

$$\begin{aligned} \frac{dQ^1_{so}(t)}{dt} &= G_1 \sum_{i=1}^{k} \dot{x}(\alpha_i)(|q_1^-(\alpha_i)| - |q_1^+(\alpha_i)|) \\ &\quad + \sum_{i=k+1}^{e} \left(\int_{x(\alpha_i)}^{\infty} G(\xi)d\xi \right) \dot{x}(\alpha_i)(|q_1^-(\alpha_i)| - |q_1^+(\alpha_i)|) \\ &\quad + \sum_{i=e+1}^{m} \left(\int_{x_f(t)}^{x(\alpha_i)} G(\xi)d\xi + \int_{x_f(t)}^{\infty} G(\xi)d\xi \right) \dot{x}(\alpha_i)(|q_1^-(\alpha_i)| - |q_1^+(\alpha_i)|) \\ &\quad + \sum_{i=k+1}^{e} \int_{x(\alpha_{i-1})}^{x(\alpha_i)} |q_1(x,t)|(-\dot{x}(q_1))G(x)dx \\ &\quad + \sum_{i=e+1}^{m+1} \int_{x(\alpha_{i-1})}^{x(\alpha_i)} |q_1(x,t)|(\dot{x}(q_1))G(x)dx \\ &\quad - 2\dot{x}_f(t)G(x_f(t)) \left(\int_{x_f(t)}^{\infty} |q_1(x,t)|dx \right) \\ &\le \mathcal{O}(1)G_1 \{\Gamma + (T.V + G_1)e(\Lambda_p)\} - \lambda_* \int_0^1 G(x)|q_1(x,t)|dx \\ &\quad + 2 \left(\frac{d}{dt} \int_{x_f(t)}^{\infty} G(\xi)d\xi \right) \int_{x_f(t)}^{\infty} |q_1(x,t)|dx, \end{aligned} \tag{4.5}$$

where we have used the fact that

$$\text{if} \quad x(q_1) < x_f(t), \quad \dot{x}(q_1) > \lambda_*, \quad \text{and if} \quad x(q_1) > x_f(t), \quad \dot{x}(q_1) < -\lambda_*.$$

Case 3 ($x_f(t) > 1$). In this case, $Q^1_{so}(t)$ becomes

$$Q^1_{so}(t) = \int_{-\infty}^{x_f(t)} |q_1(x,t)| \left(\int_{x(q_1)}^{\infty} G(\xi) d\xi \right) dx.$$

We denote the locations of waves as follows:

$$\begin{aligned} &-\infty = x(\alpha_0) < x(\alpha_1) < \cdots < x(\alpha_k) = 0 < x(\alpha_{k+1}) < \cdots < x(\alpha_l) = 1 \\ &< x(\alpha_{l+1}) < \cdots < x(\alpha_e) = x_f(t) < x(\alpha_{e+1}) < \cdots < x(\alpha_m) < x(\alpha_{m+1}) = \infty. \end{aligned}$$

Using the fact that $\int_x^\infty G(\xi) d\xi = 0, \ x \geq 1$, we have

$$\begin{aligned} \frac{dQ^1_{so}(t)}{dt} &= G_1 \sum_{i=1}^{k} \frac{d}{dt} \int_{x(\alpha_{i-1})}^{x(\alpha_i)} |q_1(x,t)| dx + \sum_{i=k+1}^{l} \frac{d}{dt} \int_{x(\alpha_{i-1})}^{x(\alpha_i)} |q_1(x,t)| \\ &\quad \cdot \left(\int_{x(q_1)}^{\infty} G(\xi) d\xi \right) dx \\ &= G_1 \sum_{i=1}^{k} \dot{x}(\alpha_i)(|q_j^-(\alpha_i)| - |q_j^+(\alpha_i)|) \\ &\quad + \sum_{i=k+1}^{l} \left(\int_{x(\alpha_i)}^{\infty} G(\xi) d\xi \right) \dot{x}(\alpha_i)(|q_1^-(\alpha_i)| - |q_1^+(\alpha_i)|) \\ &\quad + \sum_{i=k+1}^{l} \int_{x(\alpha_{i-1})}^{x(\alpha_i)} |q_1(x,t)|(-\dot{x}(q_1)) G(x) dx \end{aligned} \tag{4.6}$$

$$\leq \mathcal{O}(1) G_1 \left\{ \Gamma + (T.V + G_1) e(\Lambda_p) \right\} - \lambda_* \int_0^1 G(x) |q_1(x,t)| dx. \tag{4.7}$$

Combining all estimates (4.4)–(4.7), we have

$$\begin{aligned} \frac{dQ^1_{so}(t)}{dt} \leq & -\lambda_* \int_{-\infty}^{\infty} G(x) |q_1(x,t)| dx + 2 \left(\frac{d}{dt} \int_{x_f(t)}^{\infty} G(\xi) d\xi \right) \int_{x_f(t)}^{\infty} |q_1(x,t)| dx \\ & + \mathcal{O}(1) G_1 \left\{ \Gamma + (T.V + G_1) e(\Lambda_p) \right\}. \end{aligned}$$

This completes the proof. □

Using the lemmas in sections 2 and 3, we can obtain the following time-evolution estimates of functionals defined in section 2. Since the proof of the decay rates of the functionals in $H(t)$ given in the following lemma was given in [11, 22], we omit this proof here.

LEMMA 4.3. *Under the same assumption as in Lemma* 4.1, *we have the following time-decay estimates on the component functionals. For* $t \in I_p^2$,

1. $\frac{dL(t)}{dt} \leq \mathcal{O}(1)\Gamma + \mathcal{O}(1)(T.V + G_1) e(\Lambda_p)$,
2. $\frac{dQ_d(t)}{dt} \leq -\tilde{c}\Gamma_d + \mathcal{O}(1)(\sum_{\alpha \in \mathcal{J}} G(x(\alpha))|\alpha|) L(t) + \mathcal{O}(1)(T.V + G_1)(\Gamma + e(\Lambda_p))$,
3. $\frac{dE(t)}{dt} \leq -\tilde{c}\Gamma_s + \mathcal{O}(1)(\sum_{\alpha \in \mathcal{J}} G(x(\alpha))|\alpha|) L(t) + \mathcal{O}(1)(T.V + G_1)(\Gamma + e(\Lambda_p))$,

4. $\frac{dQ_{so}(t)}{dt} \le -\tilde{c}\Gamma_{so} + 2(\frac{d}{dt}\int_{x_f(t)}^{\infty} G(\xi)d\xi)L^1(t) + \mathcal{O}(1)G_1\{\Gamma + (T.V + G_1)e(\Lambda_p)\}$, *where* $e(\Lambda_p) = Q(\Lambda_p)+C(\Lambda_p)+(\epsilon+\delta+NsG_0)$ *and* $\tilde{c}$ *is a positive constant independent of time* t.

The following lemma gives the change of the nonlinear functional $H(t)$ after time $t = Ns$.

LEMMA 4.4. *Under the same assumption as in Lemma* 4.1*, we have*

$$H(pNs+) \le H((p-1)Ns+) + \mathcal{O}(1)\left(e(\Lambda_p) + \int_{x_f((p-1)Ns)}^{x_f(pNs)} G(\xi)d\xi\right)Ns.$$

Proof. From the definition of $H(t)$ and Lemma 4.1, for $t \in I_p^2$, we have

$$\begin{aligned}
\frac{dH(t)}{dt} &= (1 + K_1F((p-1)Ns))\frac{dL(t)}{dt} + K_2\left(\frac{dQ_d(t)}{dt} + \frac{dE(t)}{dt} + \frac{dQ_{so}(t)}{dt}\right)\\
&\le [\mathcal{O}(1)\{1 + K_1F((p-1)Ns))\} + \mathcal{O}(1)K_2(T.V + G_1) + \mathcal{O}(1)K_2G_1 - \tilde{c}K_2]\Gamma\\
&\quad + \Bigg[\mathcal{O}(1)(T.V + G_1)(1 + K_1F((p-1)Ns)) + \mathcal{O}(1)K_2(T.V + G_1)\\
&\quad + \mathcal{O}(1)K_2\left[\sum G(x(\alpha))|\alpha| + \frac{d}{dt}\int_{x_f(t)}^{\infty} G(\xi)d\xi\right]L(t)\\
&\quad + \mathcal{O}(1)K_2G_1(T.V + G_1)\Bigg]e(\Lambda_p). \qquad (4.8)
\end{aligned}$$

Since $L(t)$ is Lipschitz continuous on $((p-1)Ns, pNs)$, we have

$$L(t) \le \mathcal{O}(1)Ns + L(pNs-). \qquad (4.9)$$

By definition of $H(t)$, there is a jump across $t = pNs, p \in \{1, \dots, M\}$. Next we estimate the size of this jump.

$$\begin{aligned}
H(pNs+) - H(pNs-) &= [(1 + K_1F(pNs+))L(pNs+) + K_2(Q_d(pNs+) + E(pNs+)\\
&\quad + Q_{so}(pNs+))] - [(1 + K_1F((p-1)Ns+)))L(pNs-)\\
&\quad + K_2(Q_d(pNs-) + E(pNs-) + Q_{so}(pNs-))] = \sum_{i=1}^{5} I_i,
\end{aligned}$$

where

$$\begin{aligned}
I_1 &\equiv K_1(F(pNs+) - F((p-1)Ns+))L(pNs-),\\
I_2 &\equiv (1 + K_1F(pNs+))(L(pNs+) - L(pNs-)),\\
I_3 &\equiv K_2(Q_d(pNs+) - Q_d(pNs-)),\\
I_4 &\equiv K_2(E(pNs+) - E(pNs-)),\\
I_5 &\equiv K_2(Q_{so}(pNs+) - Q_{so}(pNs-)).
\end{aligned}$$

Since $G_{11}(x, u) \le -\lambda$, it follows from Lemma 4.1 that

$$F(pNs+) - F((p-1)Ns+) \le -\frac{1}{2}(Q(\Lambda_p) + C(\Lambda_p)) - \frac{c_0}{2}\sum_{k=(p-1)N}^{pN-1}\left|\int_{x_f(ks)}^{x_f((k+1)s)} G(x)dx\right|.$$

Therefore,

$$I_1 \le -\frac{K_1}{2}\left(Q(\Lambda_p)+C(\Lambda_p)+\frac{c_0}{2}\sum_{k=(p-1)N}^{pN-1}\left|\int_{x_f(ks)}^{x_f((k+1)s)}G(x)dx\right|\right)L(pNs-). \tag{4.10}$$

On the other hand, the difference of a wave pattern at time $t = pNs+$ and $t = pNs-$ comes from interaction, cancellation, and randomness, so we have

$$L(pNs+) - L(pNs-) \le \mathcal{O}(1)e(\Lambda_p)Ns.$$

Hence

$$I_2 \le \mathcal{O}(1)(1 + K_1F(pNs+))e(\Lambda_p)Ns. \tag{4.11}$$

By definition of $Q_d(t)$, I_3 can be estimated by considering the following two terms: one term is the product of the change of the wave strengths and the L^1 norm at $t = pNs-$, and the other term is the product of the change of the L_1 norm times the wave strengths. Therefore, for some $C_2 > 0$, we have

$$I_3 \le C_2K_2(T.V + G_1)e(\Lambda_p)Ns + C_2K_2(Q(\Lambda_p) + C(\Lambda_p))L(pNs-). \tag{4.12}$$

Similarly, we have

$$I_4 \le C_2K_2(T.V + G_1)e(\Lambda_p)Ns + C_2K_2(Q(\Lambda_p) + C(\Lambda_p))L(pNs-). \tag{4.13}$$

$$I_5 \le C_2K_2G_1e(\Lambda_p)Ns. \tag{4.14}$$

Summing up all I_k's (4.10)–(4.14), we have

$$\begin{aligned} H(pNs+) - H(pNs-) \le{}& \left(2C_2K_2 - \frac{K_1}{2}\right)(Q(\Lambda_p) + C(\Lambda_p))L(pNs-) \\ &- \frac{K_1c_0}{2}\left(\sum_{k=(p-1)N}^{pN-1}\left|\int_{x_f(ks)}^{x_f((k+1)s)}G(x)dx\right|\right)L(pNs-) \\ &+ [\mathcal{O}(1)(1 + K_1F(pNs+)) + \mathcal{O}(1)K_2(T.V + G_1) \\ &+ \mathcal{O}(1)K_2G_1]e(\Lambda_p)Ns. \end{aligned} \tag{4.15}$$

If we integrate (4.8) from $(p-1)Ns$ to pNs, then, using (4.9), we have

$$\begin{aligned} &H(pNs-) - H((p-1)Ns+) \\ &\le [\mathcal{O}(1)(1 + K_1F((p-1)Ns)) + \mathcal{O}(1)K_2(T.V + G_1) \\ &+ \mathcal{O}(1)K_2G_1 - \tilde{c}K_2]\int_{(p-1)Ns}^{pNs}\Gamma(t)dt + \mathcal{O}(1)K_2\left(\int_{x_f((p-1)Ns)}^{x_f(pNs)}G(\xi)d\xi\right)Ns \\ &+ [\mathcal{O}(1)(T.V + G_1)(1 + K_1F((p-1)Ns)) + \mathcal{O}(1)K_2(T.V + G_1) \\ &+ \mathcal{O}(1)K_2G_1(T.V + G_1)]e(\Lambda_p)Ns + \mathcal{O}(1)K_2(Q_1(\Lambda_p) + Q_2(\Lambda_p))Ns \\ &+ \mathcal{O}(1)K_2(Q_1(\Lambda_p) + Q_2(\Lambda_p))L(pNs-) \\ &+ \mathcal{O}(1)K_2\left(\int_{x_f((p-1)Ns)}^{x_f(pNs)}G(\xi)d\xi\right)L(pNs-), \end{aligned} \tag{4.16}$$

where the integral is over $((p-1)Ns, pNs)$, and we have used the fact that $\int_{(p-1)Ns}^{pNs} \sum_{\alpha\in\mathcal{J}} G(x(\alpha))|\alpha|dt = \mathcal{O}(1)(Q_1(\Lambda_p) + Q_2(\Lambda_p))$. From (4.15) and (4.16), we have

$$
\begin{aligned}
&H(pNs+) - H((p-1)Ns+)\\
&\le [\mathcal{O}(1)(1 + K_1F((p-1)Ns)) + \mathcal{O}(1)K_2(T.V + G_1)\\
&+ \mathcal{O}(1)K_2G_1 - \tilde{c}K_2] \int \Gamma(t)dt\\
&+ [\mathcal{O}(1)(T.V + G_1)(1 + K_1F((p-1)Ns)) + \mathcal{O}(1)K_2(T.V + G_1)\\
&+ \mathcal{O}(1)K_2G_1(T.V + G_1) + \mathcal{O}(1)(1 + K_1F(pNs+))\\
&+ \mathcal{O}(1)K_2G_1 + \mathcal{O}(1)K_2]e(\Lambda_p)Ns + \mathcal{O}(1)K_2 \left(\int_{x_f((p-1)Ns)}^{x_f(pNs)} G(\xi)d\xi \right) Ns\\
&+ \left[2C_2K_2 + \mathcal{O}(1)K_2 - \frac{K_1}{2} \right] (Q(\Lambda_p) + C(\Lambda_p))L(pNs-)\\
&+ \left(\mathcal{O}(1)K_2 \int_{x_f((p-1)Ns)}^{x_f(pNs)} G(\xi)d\xi - \frac{K_1c_0}{2} \sum_{k=(p-1)N}^{pN-1} \left| \int_{x_f(ks)}^{x_f((k+1)s)} G(x)dx \right| \right) L(pNs-).
\end{aligned}
$$

Since $F(t), G_0, G_1$, and $T.V$ are sufficiently small, we can choose positive constants K_1 and K_2 so that

$$
\begin{aligned}
&\mathcal{O}(1)(1 + K_1F((p-1)Ns)) + \mathcal{O}(1)K_2(T.V + G_1) + \mathcal{O}(1)K_2G_1 - \tilde{c}K_2 < 0,\\
&2C_2K_2 + \mathcal{O}(1)K_2 - \frac{K_1}{2} < 0, \quad \mathcal{O}(1)K_2 - \frac{K_1c_0}{2} < 0.
\end{aligned}
$$

Then, for such K_1 and K_2, we have

$$
H(pNs+) \le H((p-1)Ns+) + \mathcal{O}(1) \left(e(\Lambda_p) + \mathcal{O}(1) \int_{x_f((p-1)Ns)}^{x_f(pNs)} G(\xi)d\xi \right) Ns.
$$

This completes the proof. ∎

Using Lemma 4.4 successively, we obtain the following estimates.

LEMMA 4.5. *Let $v(x,t)$ be a Glimm solution corresponding to the small perturbation $v_0(x)$ of $u(x)$. Suppose the condition* (1.3) *holds. Then we have*

$$
H(T) \le H(0) + \mathcal{O}(1) \left(Q(\Lambda_T) + C(\Lambda_T) + \int_{x_f(0)}^{x_f(T)} G(\xi)d\xi \right) Ns + \mathcal{O}(1)(\epsilon + \delta + NsG_0)T.
$$

Proof. Let $v_r(x,t)$ and $u_r(x)$ be the simplified wave patterns and $T = MNs$. By Lemma 4.4, we have

$$
H(MNs+) \le H((M-1)Ns+) + \mathcal{O}(1) \left(e(\Lambda_M) + \int_{x_f((M-1)Ns)}^{x_f(MNs)} G(\xi)d\xi \right) Ns.
$$

If we use Lemma 4.4 successively in p, we obtain

$$
H(T) \le H(0) + \mathcal{O}(1) \left(Q(\Lambda_T) + C(\Lambda_T) + \int_{x_f(0)}^{x_f(T)} G(\xi)d\xi \right) Ns + \mathcal{O}(1)(\epsilon + \delta + NsG_0)T.
$$

Now we can complete the proof of Theorem 1.2 as follows. ∎

Proof of Theorem 1.2. Let $v_r(x,t)$ and $u_r(x)$ be two simplified wave patterns such that

$$\lim_{r,\epsilon,\delta\to 0} v_r(x,t) = v(x,t), \qquad \lim_{r,\epsilon,\delta\to 0} u_r(x) = u(x) \quad \text{in } L^1_{loc}(\mathbb{R}\times\mathbb{R}_+).$$

Since $H[v(\cdot,t),u(\cdot)] = \lim_{r,\epsilon,\delta\to 0} H[v_r,u_r]$, it follows from Lemma 4.4 that

$$H(t) \le H(0).$$

Since $H[v(\cdot,t),u(\cdot)]$ is equivalent to $\|v(\cdot,t)-u(\cdot)\|_{L^1(\mathbb{R})}$, i.e.,

$$\frac{1}{C_3}\|v(\cdot,t)-u(\cdot)\|_{L^1(\mathbb{R})} \le H(t) \le C_3\|v(\cdot,t)-u(\cdot)\|_{L^1(\mathbb{R})} \quad \text{for some positive constant } C_3,$$

we therefore have

$$\|v(\cdot,t)-u(\cdot)\|_{L^1(\mathbb{R})} \le C_3H(t) \le C_3H(0) \le C_3^2\|v(\cdot,0)-u(\cdot)\|_{L^1(\mathbb{R})}.$$

This completes the proof. □

Acknowledgments. The authors wish to thank Professors Alberto Bressan and Tai-Ping Liu for their insightful discussions. The authors also would like to thank the referees for their helpful comments.

REFERENCES

[1] D. Amadori and G. Guerra, *Uniqueness and continuous dependence for systems of balanced laws with dissipation*, Nonlinear Anal., 49 (2002), pp. 987–1014.
[2] D. Amadori, L. Gosse, and G. Guerra, *Global BV entropy solutions and uniqueness for hyperbolic systems of balance laws*, Arch. Ration. Mech. Anal., 162 (2002), pp. 327–366.
[3] A. Bressan and R. M. Colombo, *The semigroup generated by* 2×2 *conservation laws*, Arch. Ration. Mech. Anal., 133 (1995), pp. 1–75.
[4] A. Bressan, G. Crasta, and B. Piccoli, *Well-posedness of the Cauchy problem for* $n\times n$ *systems of conservation laws*, Mem. Amer. Math. Soc., 146 (2000).
[5] A. Bressan, T.-P. Liu, and T. Yang, L^1 *stability estimates for* $n\times n$ *conservation laws*, Arch. Ration. Mech. Anal., 149 (1999), pp. 1–22.
[6] G. Q. Chen and J. Glimm, *Global solutions to the compressible Euler equations with geometrical structure*, Comm. Math. Phys., 180 (1996), pp. 153–193.
[7] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves,* Interscience, New York, 1948.
[8] G. Crasta and B. Piccoli, *Viscosity solutions and uniqueness for systems of inhomogeneous balance laws*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 477–502.
[9] P. Embid, J. Goodman, and A. Majda, *Multiple steady states for* 1*-D transonic flow*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 21–41.
[10] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.
[11] S.-Y. Ha, L^1 *stability for systems of conservation laws with a nonresonant moving source*, SIAM J. Math. Anal., 33 (2001), pp. 411–439.
[12] A. L. Hoffman, *A single fluid model for shock formation in MHD shock tubes*, J. Plasma Phys., 1 (1967), pp. 192–207.
[13] P. D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.
[14] W.-C. Lien, *Hyperbolic conservation laws with a moving source*, Comm. Pure Appl. Math., 52 (1999), pp. 1075–1098.
[15] T.-P. Liu, *The deterministic version of the Glimm scheme*, Comm. Math. Phys., 57 (1977), pp. 135–148.
[16] T.-P. Liu, *Quasilinear hyperbolic systems*, Comm. Math. Phys., 68 (1979), pp. 141–172.
[17] T.-P. Liu, *Transonic gas flow in a duct of varying area*, Arch. Ration. Mech. Anal., 80 (1982), pp. 1–18.

[18] T.-P. Liu, *Nonlinear stability and instability of transonic gas flows through a nozzle*, Comm. Math. Phys., 83 (1982), pp. 243–260.
[19] T.-P. Liu, *Nonlinear resonance for quasilinear hyperbolic equation*, J. Math. Phys., 28 (1987), pp. 2593–2602.
[20] T.-P. Liu and T. Yang, *A new entropy functional for a scalar conservation law*, Comm. Pure Appl. Math., 52 (1999), pp. 1427–1442.
[21] T.-P. Liu and T. Yang, L^1 *stability for* 2×2 *systems of hyperbolic conservation laws*, J. Amer. Math. Soc., 12 (1999), pp. 729–774.
[22] T.-P. Liu and T. Yang, *Well posedness theory for hyperbolic conservation laws*, Comm. Pure Appl. Math., 52 (1999), pp. 1553–1586.
[23] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1982.
[24] B. Whitham, *Linear and Nonlinear Waves*, John Wiley, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1253–1278

ON THE DISCRETE SPECTRUM OF SYSTEMS IN THE PLANE AND THE DAVEY–STEWARTSON II EQUATION*

JAVIER VILLARROEL† AND MARK J. ABLOWITZ‡

Abstract. The discrete spectrum of first order systems in the plane and localized solutions of the Davey–Stewartson II equation are studied via the inverse scattering transform. Localized nonsingular algebraically decaying potentials are found which correspond to a discrete spectrum whose related eigenfunctions have, in general, multiple poles and are associated to kernels with dimension ≥ 2. There is an associated index, or winding number, which is used to classify these potentials. With suitable assumptions the mass of the corresponding Davey–Stewartson solution is found to be proportional to the index.

Key words. integrable equations, spectral analysis, interacting solitons

AMS subject classifications. Primary, 35Q58; Secondary, 35Q53, 35P25

PII. S0036141001391627

1. Introduction.

In this paper we study the classification and properties of the discrete spectrum associated with the linear differential operator in the plane (the Dirac system)

$$L\psi \equiv (\partial_x + iJ\partial_y - A)\,\psi = 0, \tag{1}$$

where J is a real constant diagonal $n \times n$ matrix, $J = \text{Diag}(J_1, \dots, J_n)$, $J_1 \neq J_2 \neq \cdots \neq J_n$, and $A(x, y)$ is an off-diagonal $n \times n$ matrix of decaying potentials.

The above problem, often referred to as the Dirac system, has attracted significant interest by itself. The Dirac system, along with the nonstationary Schrödinger (NS) equation, is one of the most relevant linear operators in two space dimensions. The importance of this problem is highlighted by noting that the solution to the Cauchy problem, corresponding to decaying data, to some of the most interesting two space, one time $(2+1)$ dimensional integrable nonlinear equations can be obtained via the inverse scattering transform (IST) and associated spectral analysis of the operator (1). Foremost among them is the Davey–Stewartson II (DSII) equation

$$iq_t + \frac{1}{2}(q_{yy} - q_{xx}) + q\,(R_{yy} - R_{xx}) = 0, \tag{2.1}$$

$$\left(\partial_y^2 + \partial_x^2\right) R(x, y) = -\sigma|q|^2, \tag{2.2}$$

which corresponds to $n = 2$ and certain reductions (see formula (18) below). Here $\sigma^2 = 1$.

Equations (2.1) and (2.2) describe, upon scaling of the physical parameters, two-dimensional quasi-monochromatic wave packets in dispersive media; here $q(x, y, t)$ is

*Received by the editors June 28, 2001; accepted for publication (in revised form) October 31, 2002; published electronically May 12, 2003. This work was partially sponsored by the Air Force Office of Scientific Research under grant F49620-00-1-0031, by NSF grant DMS-9703850, and by CICYT grant BFM2002-02609 and Junta de Castilla-Leon SA078/03 in Spain.
http://www.siam.org/journals/sima/34-6/39162.html

†Universidad de Salamanca, Facultad de Ciencias, Plaza Merced, 37008 Salamanca, Spain (javier@gugu.usal.es).

‡Department of Applied Mathematics, University of Colorado, Boulder, CO 80309-0526 (mark.ablowitz@colorado.edu).

the complex amplitude of the wave. Applications include both fluid dynamics [1, 2] and plasma physics [3]. Interesting solutions of this and other integrable equations are the lumps: localized wave configurations decaying rationally at infinity that asymptotically move with uniform velocities. (We call this an N-lump solution if the result consists of N of these objects. A class of such solutions was constructed in [4].) After these waves interact they asymptotically recover their velocity and size. This kind of behavior is usually expected for localized solutions of integrable equations. From a spectral perspective, the N-lump solution is associated with wave functions that have simple poles and hence correspond to the discrete spectrum of the spectral problem (1).

IST has been employed to solve (formally) other initial value problems for a number of important nonlinear evolution equations appearing in physics ([5, 6]; see also [7]). The relevant ideas in multidimensions regarding the continuous spectrum of the operator (1) were developed in [8], where the continuous spectrum is related to a $\bar{\partial}$ (DBAR) problem, a nontrivial generalization of the Riemann problems that appear in the one-dimensional case. We note that the IST has also been used to obtain the Hamiltonian structure and the action angle variables of (2.1), (2.2) [9, 10]. Another major difference with regard to one dimension stems from the fact that in multidimensions "small" norm assumptions are required, rendering the analysis incomplete; in particular homogeneous solutions, which make up the discrete spectrum and are intimately related to the lumps of the associated integrable equation, are beyond the scope of the theory. In terms of the rigorous theory this situation is quite unsatisfactory, as it fails to explain the most physically interesting solutions.

Recently, new localized solutions possessing nontrivial dynamics have been found for several integrable equations [11, 12, 13, 14, 15]. The connection with the discrete spectrum of the relevant spectral problem has been described in the case of the NS operator and the Kadomtsev–Petviashvili I equation (KPI) [11, 12]. This new class of solutions is found to correspond to wave functions that have higher order poles in the spectral variable. As it happened with the continuous part, the description of the discrete spectrum in multidimensions was found to involve novel features not present in one-dimensional problems.

The present work continues the development of discrete spectral theory in multidimensions, first set forth in [12] in connection with the NS operator. We find a class of rationally decaying, regular, localized potentials of the Dirac system (1) that yield meromorphic wave functions with simple or multiple poles in the spectral variable. We also consider the general case when both continuous and discrete spectrums are present. In this regard we note the following facts: (i) Different potentials exist (apparently infinitely many) that correspond to the same analytic structure (i.e., simple, double ... poles) of the wave function; this degeneracy of the discrete spectrum is explained in terms of a new topological number or index Q. (ii) The corresponding DSII solution describes the interaction of lumps with nontrivial dynamics; both (i) and (ii) hold even for eigenfunctions with simple poles and no continuous spectrum (see section 4 below). These remarkable facts have so far not been observed for any other integrable equation including the KPI equation. (iii) The results of [4] correspond to simple poles with $Q = 1$. Different values of Q and/or higher order pole multiplicities yield new DSII solutions. (iv) We obtain several different representations of the index and show that it is a winding number. (v) We prove that the mass of the lump is proportional to the index, namely (see (33)), $(4\pi)^{-1}\|q\|_2^2 = Q$, and hence it is "quantized" (can only take integer values).

Remark. 1. Ample classes of localized solutions of integrable equations can be derived using direct Darboux methods (see [13] in the context of DSII, and also [16, 17, 18] in connection with KPI equation). Despite the latter fact, direct methods cannot be used to solve the corresponding initial value problem, to study the interaction of radiation and localized solutions, or to obtain the Hamiltonian structure with the action angle variables and constants of the motion. Neither do they provide information regarding the associated and interesting linear problem; spectral analysis, in contrast, does, thus giving important and much deeper insight into the study.

2. We also find remarkable differences between the Dirac operator (DO) and the other natural spectral operator in the plane, NS. The most obvious is that the latter is a scalar operator while the former is a matrix operator. Hence so is the index Q, with several complications arising from this fact. Another important difference regards the fact that at eigenvalues the dimension d of the null space of homogeneous solutions of the DO is not restricted to be one (as happens in NS) and typically is found to be two (at least); we note that the appearance of lumps of DSII is intimately related with this possibility. Equally significant is the fact that even for the case of simple pole eigenfunctions an infinite degeneracy regarding the index (independent of that of the null space) is found, with every integer value of the index permitted; hence there exist different localized potentials associated to wave functions having simple poles. This extra freedom is inherited by the associated integrable equations. Corresponding to pure simple pole eigenfunctions the NS operator requires $Q = 1$, and hence there is only one such function. The corresponding lump of KPI decays as $\frac{1}{r^2}$ and has trivial dynamics. In contrast, for the DO the degeneracy in the dimension of the null space opens the possibility of having localized solutions with stronger decay (as $\frac{1}{r^3}, \dots$). Besides, the infinite degeneracy of the index implies that there exist localized solutions of DSII that decay as $|q|^2 = -\sigma(\partial_y^2 + \partial_x^2)\log(r^{2Q} + \dots)$ for any integer Q. Generically these solutions have a nontrivial dynamics. Another remarkable fact, not found for KPI, is the following. Given a localized solution q of DSII, the physically related state $\tilde{q}(x, y, t) \equiv \bar{q}(x, y, -t)$, obtained by backwards evolution and conjugation of phase, can correspond to a totally different spectral description and singularity structure.

The table below summarizes the differences for the case of (only) simple poles.

Operator	Type	d	Q	Equation	Dynamics of associated lumps	Decay	Number of lumps
Dirac	$n \times n$ matrix	1, 2	any integer	DSII	nontrivial	$\frac{Q}{r^2}$, any Q, or $\frac{1}{r^3}, \dots$	Q
NS	scalar	1	1	KPI	trivial	$\frac{1}{r^2}$	1

1.1. A review of known results on the spectrum of the DO. We consider here the spectral theory for the general DO where ψ is an $n \times n$ matrix. It is convenient to introduce a related function μ_{lj}, $\mu_{lj} = \psi_{lj} e^{-k(iJ_j x - y)}$ and find that μ satisfies

$$(\partial_x + iJ_l\partial_y + ik(J_j - J_l))\,\mu_{lj} - (A\mu)_{lj} = 0, \qquad l, j = 1, \dots, n. \tag{3}$$

If the potential is suitably decaying as $x^2 + y^2 \to \infty$, we can convert the above equation normalized to $\mu \to I$ as $x^2 + y^2 \to \infty$ and $|k| \to \infty$ into an integral equation

$$\mu_{lj}(\tilde{x}, k) = \delta_{lj} + \int\int G_{lj}(\tilde{x} - \tilde{x}', k)(A\mu)_{lj}(\tilde{x}', k)d\tilde{x}' \tag{4}$$

(integration is over the plane). Here $\tilde{x}$ stands for the coordinate pair (x, y) and we denote $d\tilde{x} \equiv dxdy$. Green's function is given by

$$G_{lj}(\tilde{x}, k) = \frac{\text{sign } J_l}{2\pi(J_l x + iy)} e^{i\theta_{lj}}, \tag{5}$$

where

$$\theta_{lj}(k) \equiv c_{-lj}(J_l k_R x + k_I y), \qquad c_{\pm lj} \equiv \frac{J_l \pm J_j}{J_l} \tag{6}$$

(c_{+lj} will be used later). We find it convenient also to consider these equations in columnwise form. Thus the jth column $\vec{\mu}_j, (\vec{\mu}_j)_i = \mu_{ij}$ satisfies

$$\vec{\mu}_j(\tilde{x}, k) = \vec{I}_j + \int\int \vec{G}_j(\tilde{x} - \tilde{x}', k)(A\vec{\mu})_j(\tilde{x}', k) d\tilde{x}'. \tag{7}$$

We next define the operator $\mathcal{G}_j(k)$ acting on column vectors by

$$\mathcal{G}_j(k)\vec{\mu}(\tilde{x}, k) \equiv \vec{\mu}(\tilde{x}, k) - \int\int \vec{G}_j(\tilde{x} - \tilde{x}', k)(A\vec{\mu})_j(\tilde{x}', k) d\tilde{x}' \tag{8}$$

in terms of which (7) reads $\mathcal{G}_j(k)\vec{\mu}_j = \vec{I}_j$. Note that $(\vec{I}_j)_i = \delta_{ij}$ and $(\vec{G}_j)_i = G_{ij}$.

Equation (7) implies that generically its solution $\vec{\mu}_j(\tilde{x}, k)$ is nowhere holomorphic as a function of $k \equiv k_R + ik_I$. This departure from holomorphicity defines the continuous spectrum of the problem. We recall the basic results in this regard. Define the $n \times n$ matrix Ω^{lj} by $\{\Omega^{lj}\}_{\alpha\beta} = e^{i\theta_{lj}} T_{lj} \delta_{l\alpha} \delta_{j\beta}$, where

$$T_{lj}(k) = \frac{ic_{-lj}\text{sign}J_l}{4\pi} \int e^{-i\theta_{lj}} (A\mu)_{lj} d\tilde{x} \tag{9}$$

are the "continuous" scattering data. Then one finds that

$$\frac{\partial \mu}{\partial \bar{k}} = \sum_{j,l=1}^{n} \mu \left(k_R + i\frac{J_j}{J_l} k_I \right) \Omega^{lj}(k). \tag{10}$$

Using that $\mu \to I$ as $|k| \to \infty$, the generalized Cauchy formula yields

$$\mu(k) = I + \frac{1}{2\pi i} \int_C \frac{\frac{\partial \mu}{\partial \bar{z}}}{z - k} dz \wedge d\bar{z}. \tag{11}$$

The departure from holomorphicity (10) and (11) are the basic inverse problem equations. They define a so-called $\bar{\partial}$ (DBAR) problem.

We also recall the inverse formula to reconstruct the potential:

$$A_{lj} = i(J_j - J_l)\xi_{lj}, \qquad \text{where } \mu_{lj} = \delta_{lj} + \frac{\xi_{lj}}{k} + O(1/k^2), \qquad k \to \infty. \tag{12}$$

The study of the continuous spectrum of the integral equation (4) was carried out formally in [8] evaluating the above DBAR derivative. Rigorous properties of solutions to (4) were obtained in [19] (see also [20, 21, 22]). Existence and uniqueness for the direct problem (4) are guaranteed if the potentials are in $L_\infty \cap L_1$ and "small" enough (see (24) below).

If the small norm condition is not satisfied, solutions $\vec{\omega}$ to the homogeneous equation $\mathcal{G}_j(k_1)\vec{\omega} = \vec{0}$ may exist at some points $k = k_1$ (the eigenvalues). The span (vector space) of all such functions is denoted Ker $\mathcal{G}_j(k_1)$. By the discrete spectrum of the operator (1) we mean the set $\mathcal{E}$ of all eigenvalues; we say that a potential corresponds (purely) to this part of the spectrum when there exist solutions $\vec{\omega}$ to the homogeneous equation $\mathcal{G}_j(k_1)\vec{\omega} = \vec{0}$ at some points $k = k_1 \in \mathcal{E}$ and the continuous data (9) vanishes. Note that in general $\mathcal{E} \neq \emptyset$ if the norms $\|q\|_\infty, \|q\|_1$ are only bounded. The study of the analytic properties of eigenfunctions corresponding to the latter potentials is beyond the theory developed in [19, 20, 21, 22] and has so far not been considered in detail. In this case the above description of the inverse problem is not complete; one can expect that generically $\mu(k)$ will have singularities. Indeed, suppose that the jth column $\vec{\mu}_j(k)$ of μ has a pole k_1 with multiplicity m and tends to I_j as $k \to \infty$. Around this pole $\vec{\mu}_j(k)$ has the Laurent expansion

$$\vec{\mu}_j(k) = \vec{\mu}_{j\text{sing.}}(k) + \vec{\mu}_{j\text{reg.}}(k),$$

$$\vec{\mu}_{j\text{sing.}}(k) \equiv \sum_{r=1}^{m} \frac{\vec{\Psi}_j^r}{(k-k_1)^r}, \qquad \vec{\mu}_{j\text{reg.}}(k) \equiv \sum_{r=0}^{\infty} \vec{\nu}_j^r (k-k_1)^r + \sum_{r=1}^{\infty} \vec{\zeta}_j^r (\bar{k}-\bar{k}_1)^r, \tag{13}$$

where $\vec{\mu}_{j,reg}(k)$ is regular (but in general not analytic) in the neighborhood of k_1 and tends to $\vec{I}_j$ as $k \to \infty$. Coefficients corresponding to $r = 0$ and the principal part or residue with $r = 1$ deserve particular interest, and we shall use special notation for them. Accordingly we denote

$$\vec{\Psi}_j^1 = \text{ Residue of } \vec{\mu}_j(k) \text{ at } k_1 \equiv \vec{\Phi}_j,$$

$$\vec{\nu}_j^0 = \lim_{k \to k_1} [\text{Regular part of } \vec{\mu}_j(k) \text{ at } k_1] \equiv \vec{\nu}_j. \tag{14}$$

Letting $k \to k_1$ shows that $\vec{\Psi}_j^m$ satisfies $\mathcal{G}_j(k_1)\vec{\Psi}_j^m = \vec{0}$, which means that k_1 must be an eigenvalue and $\vec{\Psi}_j^m \in$ Ker $\mathcal{G}_j$. Thus the existence of wave functions with pole singularities (singular eigenfunctions, for short) requires that $\mathcal{E} \neq \emptyset$, and hence they are naturally associated with the discrete spectrum.

In general we expect that the only singularities of $\vec{\mu}_j(k)$ are poles of any order in k. This can be substantiated as follows; first, we note that $\mathcal{L}_j \equiv \mathcal{G}_j - I$ is a Fredholm operator [20]. The case with compactly supported potentials is particularly easy to understand, as polar operators (i.e., with the kernel having only weak singularities and vanishing away a bounded set) are well known to be Fredholm. Potentials supported on the entire plane can be viewed as having compact support on the compactified plane; the result then follows, noting that $\sup_{\mu, \|\mu\|_\infty \leq 1} \mathcal{L}_j(k)\vec{\mu}(\tilde{x}, k)$ goes to zero uniformly as $r^2 \equiv x^2 + y^2 \to \infty$ on a bounded set of functions μ. Hence Fredholm theory (and the results of [6, 8]) indicates that generically the solutions $\mu(k)$ to (4) are not holomorphic anywhere and that they may have a denumerable set of poles of any order in $k, \bar{k}$ as singularities in the finite plane. Note also that, in principle, pole singularities in the variable $\bar{k}$ are not allowed, as can be seen by expanding $\mu(k)$ for large values of $|k|$ and using (3) (we detail this in the appendix). A class of special solutions exists, however, that are analytic everywhere except at (isolated) poles; i.e., they are meromorphic. This class corresponds purely to the discrete spectrum and has localized associated potentials.

Another critical property involving the distribution of eigenvalues is the following.

Result 0. Assume that $\vec{\omega}_l$ solves the lth homogeneous equation at a point $k_l \equiv a_l + ib_l$: $\mathcal{G}_l(k_l)\vec{\omega}_l = \vec{0}$. Then $\vec{\pi}_j$ defined as $\vec{\pi}_j = e^{-i\theta_{jl}(k_l)}\vec{\omega}_l$ solves the jth homogeneous equation at the point $k_{jl} \equiv a_l + i\frac{J_l}{J_j}b_l$. Note also that $\theta_{jl}(k_l) = -\theta_{lj}(k_{jl})$.

For a proof of this remarkable fact, note that [8]

$$\omega_{il}(\tilde{x}) = \int\int G_{il}(\tilde{x} - \tilde{x}', k_l)(A\omega)_{il}(\tilde{x}')d\tilde{x}',$$

and hence

$$\pi_{ij}(\tilde{x}) = \int\int \left(e^{-i\theta_{jl}}G_{il}\right)(\tilde{x} - \tilde{x}', k_l)(A\pi)_{ij}(\tilde{x}')d\tilde{x}'.$$

The proof is finished by noting that the Green's function (equation (5)) satisfies the following symmetric relationship:

$$G_{ij}(k_{jl}) = e^{-i\theta_{jl}(k_l)}G_{il}(k_l). \tag{15}$$

In particular, taking $l = 1, j = 2, J_1 = -J_2 = 1$ we obtain that if there exists a solution $\vec{\omega}_1$ to the first homogeneous equation at a point $k_1 \equiv a + ib$, then $\vec{\pi}_2 \equiv e^{-2i(by-ax)}\vec{\omega}_1$ solves the second homogeneous equation at $\bar{k}_1$; i.e.,

$$\vec{\pi}_2 \in \text{ Ker } \mathcal{G}_2(\bar{k}_1) : \ \mathcal{G}_2(\bar{k}_1)\vec{\pi}_2 = \vec{0}. \tag{16}$$

1.2. The DSII reduction. The physical DSII problem (2.1), (2.2) is obtained with $n = 2$ and

$$J = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & q \\ r & 0 \end{pmatrix}, \quad r(\tilde{x}) = \sigma\bar{q}(\tilde{x}), \quad \tilde{x} \equiv (x, y), \quad \sigma = \pm 1, \tag{17}$$

and $\bar{q}$ stands for the complex conjugate of q. This entails a number of restrictions or symmetry relations on the wave function that we now discuss. One finds that

$$\begin{pmatrix} \mu_{12}(k) \\ \mu_{22}(k) \end{pmatrix} = \begin{pmatrix} \sigma\bar{\mu}_{21}(\bar{k}) \\ \bar{\mu}_{11}(\bar{k}) \end{pmatrix}, \tag{18}$$

and this implies in particular that the position, number, and multiplicities of the poles of different columns are related. One has the following:

(i) Singular functions μ have the structure (13) and (18), where the location of the poles of $\vec{\mu}_2$ are the complex conjugates of those of $\vec{\mu}_1$ and have the same order m. Hence the discrete spectrum is an even-dimensional set $\{k_j, \bar{k}_j\}_{j=1,\dots,N}$.

(ii) The Laurent coefficients of the first and second columns of μ,

$$\vec{\Psi}_1^r = \begin{pmatrix} \Psi_{11}^r \\ \Psi_{21}^r \end{pmatrix}, \quad \vec{\Psi}_2^r = \begin{pmatrix} \Psi_{12}^r \\ \Psi_{22}^r \end{pmatrix}, \quad r = 1, \dots, m,$$

are related as follows:

$$\Psi_{12}^r = \sigma\bar{\Psi}_{21}^r, \qquad \Psi_{22}^r = \bar{\Psi}_{11}^r. \tag{19}$$

We can thus drop the subscripts $1, 2$ of the Laurent coefficients and simply write $\vec{\Psi}_1^r \equiv \vec{\Psi}^r, \vec{\Psi}_2^r = \tau\vec{\Psi}^r$ where we define the following involution:

$$\vec{\Psi} \equiv \begin{pmatrix} \Psi_1 \\ \Psi_2 \end{pmatrix} \Rightarrow \tau\vec{\Psi} \equiv \begin{pmatrix} \sigma\bar{\Psi}_2 \\ \bar{\Psi}_1 \end{pmatrix}. \tag{20}$$

Besides, $\vec{\mu}_{2reg}(k) = \tau\vec{\mu}_{1reg}(\bar{k})$.

(iii) At each eigenvalue $k_1 \equiv a + ib, \bar{k}_1$, there are two eigenfunctions, and hence the null space is (at least) two-dimensional. One has that

$$\{\vec{\Psi}^m, e^{2i(by-ax)}\tau\vec{\Psi}^m\} \subset \text{ Ker } \mathcal{G}_1(k_1), \qquad \{\tau\vec{\Psi}^m, e^{-2i(by-ax)}\vec{\Psi}^m\} \subset \text{Ker } \mathcal{G}_2(\bar{k}_1). \tag{21}$$

(We use the notation $\{f, g\}$ to symbolize the vector span of the two function f, g.)

Indeed, according to the comments under equation (14) $\vec{\Psi}^m \in$ Ker $\mathcal{G}_1(k_1)$, $\tau\vec{\Psi}^m \in$ Ker $\mathcal{G}_2(\bar{k}_1)$. With $\omega_1 \equiv \vec{\Psi}^m$ (16) yields that $\vec{\pi}_2 = e^{-2i(by-ax)}\vec{\Psi}^m \in$ Ker $\mathcal{G}_2(\bar{k}_1)$, whereby one has that $\{\tau\vec{\Psi}^m, e^{-2i(by-ax)}\vec{\Psi}^m\} \subset$ Ker $\mathcal{G}_2(\bar{k}_1)$. Likewise we obtain ($l = 2, j = 1, \omega_2 \equiv \tau\vec{\Psi}^m$) that $\pi_1 \equiv e^{2i(by-ax)}\tau\vec{\Psi}^m \in$ Ker $\mathcal{G}_1(k_1)$.

(iv) For wave functions with simple poles $k_1, \bar{k}_1$ and residue $\vec{\Phi}$ at k_1,

$$\vec{\mu}_1(k) = \vec{\mu}_{1,reg}(k) + \frac{\vec{\Phi}}{(k - k_1)}, \tag{22}$$

the following relationship for the Laurent coefficients $\vec{\nu} \equiv \vec{\nu}^0$ and $\vec{\Phi}$ can be derived:

$$\vec{\nu} = (z + \gamma)\vec{\Phi} + \rho_1 e^{2i(by-ax)}\tau\vec{\Phi} \tag{23}$$

(where $z \equiv y - iJ_1x \equiv y - ix$, and γ and ρ_1 are constants). Corresponding to $\rho_1 = 0$, (23) was first derived in [8], noticing that $\vec{\nu} - z\vec{\Phi} \in$ Ker $\mathcal{G}_1(k_1)$, and was employed to obtain a class of solutions to DSII which decay weakly (as $1/r$, $r \equiv (x^2 + y^2)^{\frac{1}{2}}$) but are singular. The general formula (23) with $\rho_1 \neq 0$ was first derived in [4] and used to find nonsingular rational solutions that decay as $1/r^2$ at infinity (in [23] the stability of these configurations is discussed). These solutions are usually referred to as the lumps of DSII. We recover them in section 4 below.

(v) Corresponding to the focusing case ($\sigma = -1$) of DSII, convergence via iteration requires that $q(x, y)$ satisfies the condition $\|q\|_\infty\|q\|_1 < \frac{\pi}{2}$ [19]. If in addition q satisfies

$$\frac{2}{\pi}\|q\|_\infty\|q\|_1 \leq \frac{1}{(1-\tau)^2}\|\hat{q}\|_\infty\|\hat{q}\|_1 < 1, \qquad \tau^2 \equiv \frac{1}{(2\pi)^3}\|q\|_1\|\hat{q}\|_1, \tag{24}$$

where $\hat{q}$ is the Fourier transform, then it is proven that the inverse problem also has a unique solution, and hence so does the Cauchy problem for (2.1), (2.2).

(vi) The inverse formula to reconstruct the potential (12) yields in this case that

$$q \equiv A_{12} = -2i\xi_{12}, \qquad \text{where } \xi_{lj} = \lim_{k\to\infty} k\mu_{lj}, \tag{25.1}$$

$$|q|^2 = -4\sigma\frac{\partial}{\partial\bar{z}}\xi_{11}, \qquad R = \partial_z^{-1}\xi_{11}, \qquad \text{where } z \equiv y - ix. \tag{25.2}$$

2. New results for the discrete spectrum of the DO. We now elaborate on our results for the discrete spectrum of the operator (1). As has been remarked, the appearance of lumps of DSII is intimately related to the possibility of having a higher-dimensional null space Ker $\mathcal{G}_1(k_1)$, i.e., with (21). We shall see that there is also additional freedom in both (i) the order of the poles and (ii) the value of an integer-valued quantity that we call indices or charges (as we shall see they are winding numbers), which we now introduce.

At any pole $k_1 \equiv a + ib$, of the jth column $\vec{\mu}_j(k)$ we define the index of the pole as the matrix functional $Q_{lj}[\Phi, k_1] \equiv Q_{lj},\ l, j = 1, \dots, n$,

$$Q_{jj} = \frac{\text{sign } J_j}{2\pi i} \int\int (A\Phi)_{jj}(\tilde{x}')d\tilde{x}',$$

$$Q_{lj} \equiv \frac{c_{+lj}}{4\pi i} \text{ sign } J_j \int\int e^{-i\theta_{lj}(k_1)}(A\Phi)_{lj}(\tilde{x}')d\tilde{x}', \tag{26}$$

where $\vec{\Phi}_j(k)$ is the residue of $\vec{\mu}_j(k)$ at the pole k_1, the quantities c_{lj+}, θ_{lj} are defined in (6), and we recall our notation $(\vec{\mu}_j)_i = \mu_{ij}$. We shall see that meromorphic wave functions with simple or multiple poles yield a new class of rationally decaying, regular and localized potentials of (1) and (2). We find that different potentials exist that correspond to the same analytic structure of the wave function, in particular to simple poles. The degeneracy of the spectrum is classified in terms of the index. These potentials—even in the case of simple poles in the wave function—are found to correspond to values of the index that satisfy $Q_{lj} = (Q\delta_{lj})$, with Q being a positive integer. We note that both in [8] and [4] the relationship (23) and the lump potentials correspond to assuming $Q = 1$. (This has been checked a posteriori by direct evaluation of the integral (26) in [23].) In this regard, a natural problem is to determine the most general values the index may take. We can state the following.

Result 1. (i) The index can be represented as

$$Q_{lj} = \frac{c_{+lj}}{4\pi i} \text{ sign } J_j \oint_\Gamma \tilde{\Phi}_{lj}(\tilde{x})dz, \tag{27}$$

where Γ is a closed contour at infinity that winds the origin once in the positive (counterclockwise) direction and $z \equiv y - iJ_l x$, and we write $e^{-i\theta_{lj}(k_1)}\Phi_{lj} \equiv \tilde{\Phi}_{lj}$.

(ii) If $\tilde{\Phi}_{lj}$ is nonsingular, decaying with a power series expansion in $z, \bar{z}$ at infinity (see below), one has for the index matrix

$$Q_{lj} = \begin{cases} \lambda^{\frac{c_{+lj}}{2}} \text{ sign } J_j & \text{if } \tilde{\Phi}_{lj} = \frac{\lambda}{z} \text{ as } r^2 \equiv x^2 + y^2 \to \infty, \\ 0 \text{ otherwise.} \end{cases} \tag{28}$$

It will turn out that Q_{jj} are integers.

(iii) The index matrix must be diagonal:

$$Q_{lj} = 0, \qquad l \neq j. \tag{29}$$

(iv) For the Dirac system corresponding to the DSII reduction, $Q_{22} = \bar{Q}_{11}$.

(v) If for some positive integer Q is $\Phi_{jj} = \frac{d}{dz} \ln H(z, \bar{z})$ with $H(z, \bar{z}) > 0$ and $H \sim (z\bar{z})^Q$ as $r^2 \to \infty$, then

$$Q_{jj} \text{sign } J_j = Q = \text{winding number of } H. \tag{30}$$

Proof. (i) Considering (3) for Φ_{lj} and using Green's theorem, we obtain

$$\begin{aligned} \int\int e^{-i\theta_{lj}(k_1)}(A\Phi)_{lj}(\tilde{x}')d\tilde{x}' &= \int\int (\partial_x + iJ_l\partial_y)\left(e^{-i\theta_{lj}(k_1)}\Phi_{lj}\right)d\tilde{x}' \\ &= \oint_\Gamma e^{-i\theta_{lj}(k_1)}\Phi_{lj}(z, \bar{z})dz. \end{aligned}$$

(ii) Assume that at infinity $\tilde{\Phi}_{lj}(z,\bar{z})$ has the power series expansion

$$\tilde{\Phi}_{lj}(z,\bar{z}) = \sum_{n,m=0}^{\infty} \frac{\lambda_{n,m}}{z^n \bar{z}^m}, \qquad n+m \geq 1.$$

Recalling that for a large contour Γ at infinity

$$\oint_\Gamma \frac{1}{2\pi i z^n \bar{z}^m} dz = \begin{cases} 1, & n = 1-m = 1, \\ 0 & \text{otherwise}, \end{cases}$$

(27) yields that

$$Q_{lj} = \frac{c_{+lj}}{4\pi i} \text{ sign } J_j \oint_\Gamma \tilde{\Phi}_{lj}(z,\bar{z}) dz = \frac{c_{+lj}}{2} (\text{ sign } J_j) \lambda_{1,0}.$$

(iii) We skip the proof here.
(iv) This follows directly from (19): $\Phi_{22} = \bar{\Phi}_{11}$ and (27).
(v)

$$Q_j = \frac{\text{sign } J_j}{2\pi i} \oint_\Gamma \Phi_{jj}(z,\bar{z}) dz = \frac{\text{sign } J_j}{2\pi i} \oint_\Gamma \frac{d \ln H}{dz} dz = Q.$$

Note. By (ii) $Q_{lj} \neq 0$ iff at infinity $\tilde{\Phi}_{lj}$ decays like $\frac{1}{z}$. According to (iii) this holds only for the diagonal elements of the index matrix. It follows that we can simply write $Q_j \equiv Q_{jj}$ for the diagonal elements.

Conjecture. For any potential of the DSII reduction for which μ is singular, the index matrix must be an integer multiple of the identity matrix:

$$Q_{lj} = Q\delta_{lj} \text{ or } Q_{11} = Q_{22} \equiv Q \text{ a positive integer}, \qquad Q_{12} = Q_{21} = 0, \tag{31}$$

and in addition $\Phi_{11} = \frac{d}{dz} \ln H(z,\bar{z})$ with $H \sim (z\bar{z})^Q$ as $r^2 \to \infty$.

We note that in all the examples of lump-type solutions considered, this turns out to be the case; we also note that the proofs corresponding to section 4 below apply for generic singular eigenfunctions, irrespective of whether a continuous spectrum is present or not.

Index and L_2 norms. The index has been originally defined (cf. (26) and (27)) as an integral of certain functions that are inherent to the spectral space: the residues Φ_j. A natural question arises as to whether Q is directly related to the potential $q(x,y)$ itself. We see next that under certain conditions the answer is affirmative. Let

$$\|q\|_2^2 \equiv \int\int |q(x,y)|^2 dxdy$$

be the (square of the) L_2 norm. With time present (e.g., Davey–Stewartson) physically it represents the mass of the wave $q(x,y,t)$ and it is conserved in time (i.e., it is an integral of the motion). We can state the following.

Result 2. (i) Assume that $R(x,y) = \ln \Delta(x,y)$, where $\Delta(x,y) > 0$ satisfies $\Delta = (z\bar{z})^s + O((z\bar{z})^{s-1})$ as $r^2 \to \infty$ for some positive integer s and $z \equiv y - ix$ with $J_1 = 1 = -J_2 = -\sigma$. Then

$$\|q\|_2^2 = 4\pi s. \tag{32}$$

(ii) Assume $\vec{\mu}_1$ is purely meromorphic with just one pole and residue Φ_{11} satisfying $\Phi_{11} = \frac{d}{dz} \ln H(z, \bar{z})$ with $H > 0$ and $H \sim (z\bar{z})^Q$ as $r^2 \to \infty$. Then

$$4\pi Q = \|q\|_2^2. \tag{33}$$

Proof. (i) Using (2.2) with $\sigma = -1$ and Green's theorem, one has that

$$\begin{aligned}\int\int |q(x,y,t)|^2 dxdy &= \int\int \left(\partial_y^2 + \partial_x^2\right) R dxdy = \oint_\Gamma \left(dy\partial_x - dx\partial_y\right) R(\tilde{x}) \\ &= \frac{2}{i} \oint_\Gamma \frac{d}{dz} R(z,\bar{z}) dz = \frac{2}{i} \oint_\Gamma \frac{d}{dz} \ln \Delta(z,\bar{z}) dz,\end{aligned}$$

where Γ is a closed contour at infinity that winds the origin once in the positive direction. We have $R = \ln \Delta$, where we write $\Delta = \Delta_1.\Delta_2.\Delta_3$, $\Delta_1 \equiv z^s, \Delta_2 \equiv \bar{z}^s$, and $\Delta_3 \equiv 1 + O(1/(z\bar{z}))$. The remainder decays, and it is zero when integrated out along Γ. It follows by the argument principle that

$$\frac{1}{4\pi} \int\int |q(x,y,t)|^2 dxdy = \frac{1}{2\pi i} \oint_\Gamma \frac{d}{dz} \log \Delta_1(z) dz = s.$$

(ii) In this case in (25.1) $\xi_{11} = \Phi_{11}$ and by (25.2) $R(x,y) = \ln H(z,\bar{z})$; hence by (30) Q equals the index. The result follows by (i) with $\Delta(x,y) = H$ and $s = Q$.

Remarks. 1. This formula extends to the case in which μ is purely meromorphic with N simple poles $k_1, \dots, k_N$, residues $\Phi^{j)}, j = 1, \dots, N$, and different indices $Q[\Phi^{j)}, k_j], j = 1, \dots, N$. In this case $s = \sum_{j=1}^N Q[\Phi^{j)}, k_j] = \frac{1}{4\pi}\|q\|_2^2$.

2. By means of Green's theorem as well as the argument principle, we have given three different representations, (26), (27), and (33), of the index. We have also shown that it is a winding number (the reader may consult [24] in connection with the above ideas). The mass of the wave is proportional to the index for basic lump solutions of DSII, and to the sum of indices for N-lump solution. For generic potentials (solutions of DSII) we do not expect that the mass of the wave be finite.

3. Determination of the wave functions of the discrete spectrum. We shall see that the determination of singular eigenfunctions associated with the discrete spectrum requires the following integers: the number of poles N, and at every pole $m \equiv$ multiplicity of the pole, $Q \equiv$ index matrix, and $d \equiv \text{Dim Ker } \mathcal{G}_1(k_1)$. Note: we find that certain constants determine Ker $\mathcal{G}_1(k_1)$. If $\tilde{d}$ of these constants vanish, we consider that the effective dimension of the null space is $\geq d - \tilde{d}$. Given this information we find a linear relationship between a subset of the Laurent coefficients $\{\vec{\Psi}_j^r, \vec{\nu}_j^r, r = 0, 1, \dots\}$ of $\mu(k)$, which fixes the function μ. We consider here the DSII reduction. The general case is taken up later.

Simple poles. We consider first the case of simple poles. In this case we show that the formula (23) is not the only possible one corresponding to this pole structure; this implies that there exist different localized potentials of DSII that correspond to wave functions having simple poles.

Result 3. Assume that there exists a solution to (7) such that its columns $\vec{\mu}_1(k), \vec{\mu}_2(k)$ have simple poles $k_1 \equiv a + ib, \bar{k}_1$ and residues Φ_1, Φ_2, respectively. Define

$$F_{j;1} \equiv (z_j + \gamma_j), \qquad F_{j;2} \equiv \frac{1}{2}(F_{j;1}^2 + \delta_j), \qquad j = 1, 2, \tag{34}$$

where γ_j, δ_j are certain constants (under evolution γ_j, δ_j are functions of time), and we recall that $z_j \equiv y - ixJ_j$ and that (see (16), (21) for the DSII reduction) we have as homogeneous solutions

$$\vec{\Phi}_1; \pi_1 \equiv e^{2i(by-ax)}\tau\vec{\Phi}_1 \qquad \text{and} \qquad \vec{\Phi}_2 = \tau\vec{\Phi}_1, \vec{\pi}_2 = e^{-2i(by-ax)}\vec{\Phi}. \tag{35}$$

We also introduce $f_{j;r} \equiv F_{j;r}(\gamma = \delta = 0)$. The conditions (18), (19) require that $\gamma_2 = \bar{\gamma}_1, \delta_2 = \bar{\delta}_1$, which implies $F_{2;1} = \bar{F}_{1;1}, F_{2;2} = \bar{F}_{1;2}$. It follows that we can drop the column indices of these functions and write $F_{1;n} \equiv F_n, \gamma_1 \equiv \gamma, \delta_1 \equiv \delta$. Define also $\theta \equiv \theta_{21} = 2(by - ax)$. The following hold:

(i) The index matrix must be diagonal: $(Q)_{lj} = \text{ Diag } (Q_1, \bar{Q}_1)$.

(ii) If $Q_1 = 1$, then the Laurent coefficients $\vec{\nu}_j, \vec{\Phi}_j$ satisfy $\vec{\nu}_j = F_j\vec{\Phi}_j + \rho_j\vec{\pi}_j$, where ρ_j are certain constants (we take here $\rho_1 \equiv \bar{\rho}, \rho_2 \equiv \sigma\bar{\rho}_1$), or

$$\vec{\nu}_1 = F_1\vec{\Phi}_1 + \bar{\rho}e^{i\theta}\vec{\Phi}_2, \tag{36.1.1}$$

$$\vec{\nu}_2 = \bar{F}_1\vec{\Phi}_2 + \sigma\rho e^{-i\theta}\vec{\Phi}_1. \tag{36.1.2}$$

(iii) If $Q_1 = 2$, then the Laurent coefficients $\vec{\nu}_j^1, \vec{\nu}_j, \vec{\Phi}_j$ satisfy

$$\vec{\nu}_1^1 = F_1\vec{\nu}_1 + \bar{\rho}e^{i\theta}\vec{\nu}_2 - F_2\vec{\Phi}_1 + e^{i\theta}\bar{H}\vec{\Phi}_2, \tag{36.2.1}$$

$$\vec{\nu}_2^1 = \bar{F}_1\vec{\nu}_2 + \sigma\rho e^{-i\theta}\vec{\nu}_1 - \bar{F}_2\vec{\Phi}_2 + \sigma e^{-i\theta}H\vec{\Phi}_1, \tag{36.2.2}$$

where $H \equiv \rho' - \rho\tilde{F}_1$, $\tilde{F}_1 \equiv (z_1 + \gamma_1')$, and where γ_1', ρ' are certain constants (which become functions of time under evolution).

(iv) If Q_1 is any positive integer, then the Laurent coefficients $\vec{\nu}_j^{Q-1}, \dots, \vec{\nu}_j, \vec{\Phi}_j$ satisfy

$$\vec{\nu}_1^{Q-1} = \sum_{l=-1}^{Q-2}(-1)^{l+1}F_{Q-l-1}\vec{\nu}_1^l + e^{i\theta}\sum_{l=-1}^{Q-2}(-1)^{l+1}\tilde{F}_{Q-l-2}\vec{\nu}_2^l, \tag{36.Q.1}$$

$$\vec{\nu}_2^{Q-1} = \sum_{l=-1}^{Q-2}(-1)^{l+1}\bar{F}_{Q-l-1}\vec{\nu}_2^l + \sigma e^{-i\theta}\sum_{l=-1}^{Q-2}(-1)^{l+1}\tilde{F}_{Q-l-2}^*\vec{\nu}_1^l, \tag{36.Q.2}$$

where we define $F_l, \tilde{F}_l$ recursively via $\partial_{z_1}F_l + F_{l-1} = 0, F_1 = z_1$, $\partial_{z_1}\tilde{F}_l + \tilde{F}_{l-1} = 0$, $\tilde{F}_0 = \bar{\rho}$, and $\vec{\nu}_1^{-1} \equiv \vec{\Phi}_1, \vec{\nu}_2^{-1} \equiv \vec{\Phi}_2$. In addition, by obvious reasons we use $F^* \equiv \bar{F}$.

We skip here the rather technical proof.

Double poles. We briefly mention how the above theory generalizes to deal with higher order singular wave functions, in particular those for the case of poles of order 2. Assume that around $k = k_j$, $\vec{\mu}_j(k)$ has the singular representation (13) with $m = 2$, i.e.,

$$\vec{\mu}_{j\text{sing.}}(k) = \frac{\vec{\Psi}_j}{(k-k_j)^2} + \frac{\vec{\Phi}_j}{k-k_j}, \qquad j = 1, 2, \tag{37}$$

and $k_2 = \bar{k}_1$.

Note that in this case (19) implies the following relationship for the principal Laurent coefficients of the first and second columns of μ:

$$\vec{\Psi}_1 \equiv \vec{\Psi}, \quad \vec{\Psi}_2 = \tau\vec{\Psi}; \qquad \vec{\Phi}_1 \equiv \vec{\Phi}, \quad \vec{\Phi}_2 = \tau\vec{\Phi}.$$

In addition, if $\vec{\pi}_1 \equiv e^{2i(by-ax)}\vec{\Psi}_2, \vec{\pi}_2 = e^{-2i(by-ax)}\vec{\Psi}_1$, then (21) implies

$$\{\vec{\Psi}_1, \vec{\pi}_1\} \subset \text{ Ker } \mathcal{G}_1(k_1), \qquad \{\vec{\Psi}_2, \vec{\pi}_2\} \subset \text{ Ker } \mathcal{G}_2(\bar{k}_1).$$

Define also (similar to the case of simple poles) the functions (see (34))

$$F_{j;1} \equiv (z_j + \gamma_j); \quad F_{j;2} \equiv \frac{1}{2}(F_{j;1}^2 + \delta_j), \quad H_j \equiv \rho_j' - \rho_j\tilde{F}_j; \quad \tilde{F}_j \equiv (z_j + \gamma_j'), \tag{38}$$

where $\gamma_j, \delta_j, \rho_j, \gamma_j', \rho_j'$ are constants (functions of time under evolution).

Result 4. Assume that there exists a solution to (7) such that its columns $\vec{\mu}_1(k), \vec{\mu}_2(k)$ have double poles. Then the charge $Q_j(\Phi, k_1)$ takes integer values with $Q_j \geq 2$. If $Q_j(\Phi, k_1) = 2, Q_j(\Psi, k_1) = 0$, then the Laurent coefficients $\vec{\nu}_j, \vec{\Phi}_j, \vec{\Psi}_j$ are linearly related as follows:

$$\vec{\Phi}_j = F_{j;1}\vec{\Psi}_j + \rho_j\pi_j, \tag{39.1}$$

$$\vec{\nu}_j = F_{j;2}\vec{\Psi}_j - H_j\vec{\pi}_j. \tag{39.2}$$

4. DSII potentials. The former development has a direct bearing on the construction of localized, nonsingular rationally decaying solutions of the DSII equation. Upon considering particular analytic structures for the wave functions and relevant indices, using the inverse formula (25) we can reconstruct the potential, with the appropriate temporal dependence of the scattering data (the constants $(\gamma, \delta, \dots)$) inserted.

Recall that the DSII equations (2.1), (2.2) arise from the compatibility $[L, M] = 0$, with L given by (1) and with $M = -\partial_t + A_1 - A\partial_y + iJ\partial_{yy}$. Here A_1 is given by

$$A_1 = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \tag{40}$$

with elements satisfying

$$(\partial_x + i\partial_y)A_{11} = i\frac{\sigma}{2}(\partial_x - i\partial_y)|q|^2, \qquad (\partial_x - i\partial_y)A_{22} = -i\frac{\sigma}{2}(\partial_x + i\partial_y)|q|^2,$$

$$A_{12} = -\frac{i}{2}(\partial_x - i\partial_y)q, \qquad a_{21} = \frac{\sigma}{2}(i\partial_x - \partial_y)\bar{q}. \tag{41}$$

The temporal evolution of the scattering data follows in the standard way by substituting the relevant eigenfunctions in the time operator evolution M. The constants $\gamma_j, \delta_j, \dots$ of section 3 are found to satisfy the following:

Simple poles.

$$\partial_t\gamma_j = -2iJ_jk_j, \quad \partial_t\delta_j = 2iJ_j, \quad \partial_t\rho = 2i(a^2 - b^2)\rho, \quad j = 1, 2, \tag{42}$$

and hence the temporal evolution is given by

$$\gamma_j(t) = \gamma_j(0) - 2iJ_jk_jt, \quad \delta_j(t) = \delta_j(0) + 2iJ_jt, \quad \rho(t) = \rho(0)\exp 2i(a^2 - b^2)t \tag{43}$$

both for quantities with or without primes.

Double poles. In this case one has that the temporal evolution is given by

$$\gamma_j(t) = \gamma_j(0) - 2iJ_jk_jt, \qquad \rho(t) = \rho_0 e^{2i(a^2-b^2)t} \tag{44.1}$$

both for quantities with or without primes, and (notice the change in sign)

$$\delta_j(t) = \delta_j(0) - 2iJ_jt. \tag{44.2}$$

We consider here the simplest examples of DSII solutions that correspond to pure discrete spectrums; in this case the wave function is purely meromorphic.

Simple poles. Assume that the column components of wave functions $\vec{\mu}_1(k), \vec{\mu}_2(k)$ have simple poles $k_1 \equiv a + ib$ and $\bar{k}_1$, residues $\vec{\Phi}_1, \vec{\Phi}_2$, respectively, and $\vec{\mu}_{1,2reg} = \vec{I}_{1,2}$; i.e., take $N = m = 1$ and

$$\vec{\mu}_1(k) = \vec{I}_1 + \frac{\vec{\Phi}}{(k-k_1)}, \qquad \vec{\mu}_2(k) = \vec{I}_2 + \frac{\vec{\Phi}_2}{(k-\bar{k}_1)}. \tag{45}$$

We obtain the following:

(i) $Q = 1$. Then the relevant system is (36.1) with $\vec{\nu}_j(k) = \vec{I}_j$. Upon solution and use of (25) we find that the eigenfunctions and potentials are given by [4]

$$\vec{\Phi}_2 = \tau\vec{\Phi}_1 = \frac{1}{\Delta}\begin{pmatrix} \sigma\rho e^{-i\theta} \\ F_1 \end{pmatrix}, \tag{46}$$

$$q(x,y,t) = -2i\rho\sigma\frac{e^{-i\theta}}{\Delta}, \qquad R = \ln\Delta, \tag{47}$$

$$\Delta \equiv |F_1|^2 - \sigma|\rho|^2 \equiv (x - \gamma_I + 2at)^2 + (y + \gamma_R + 2bt)^2 - \sigma|\rho|^2, \tag{48}$$

where $\theta \equiv 2(by - ax + (b^2 - a^2)t)$. This solution is characterized by the integers $m = N = Q = 1$ and $d = 2$.

The potential takes a simpler form upon transformation to a frame moving with the wave with coordinates

$$\hat{x} = x - \gamma_I + 2at, \qquad \hat{y} = y + \gamma_R + 2bt. \tag{49}$$

Indeed, relative to this coordinate frame

$$q = -2i\sigma\rho\frac{e^{2i(a\hat{x}-b\hat{y}+(b^2-a^2)t+a\gamma_I+b\gamma_R)}}{\hat{y}^2 + \hat{x}^2 - \sigma|\rho|^2}. \tag{50}$$

Note that upon transformation of the Galilean frame, the field $q(x,y,t)$ is not invariant but picks a phase factor $4(b^2 - a^2)t + 2a\gamma_I + 2b\gamma_R$.

Corresponding to $\sigma = -1$ and taking $\rho \neq 0$ the solution is nonsingular and decays rationally as $1/(x^2+y^2)$ at infinity. Hence it has all the $L_p, p > 1$ norms finite. Indeed by direct calculation one finds

$$\|q\|_1 = \infty, \quad \|q\|_p \equiv \left(\int\int |q(x,y,t)|^p dxdy\right)^{\frac{1}{p}} = 2\left(\frac{\rho^{2-p}\pi}{p-1}\right)^{\frac{1}{p}}, \quad p > 1. \tag{51}$$

This implies that all the constants of the motion (see [9]) are finite. The opposite situation arises in the defocusing case ($\sigma = 1$), where the solution is singular and hence the $L_p, p > 1$ norms are all infinite (unlike what was claimed in [4]).

From a physical perspective the solution for $\sigma = -1$ describes a wave (lump) traveling with constant velocity $v_x = -2a, v_y = -2b$, and amplitude $|q| = \frac{2}{|\rho|}$ modulated by the phase $\theta(x, y, t)$. However, this wave is unstable against perturbations [23].

(ii) If $\mu(k)$ has the structure (45) and $Q = 2$, then the Laurent coefficients satisfy (36.2) with $\vec{\nu}_j^1 = 0$:

$$F_1\vec{I}_1 + \bar{\rho}e^{i\theta}\vec{I}_2 - F_2\vec{\Phi}_1 + e^{i\theta}\bar{H}\vec{\Phi}_2 = 0,$$

$$\bar{F}_1\vec{I}_2 + \sigma\rho e^{-i\theta}\vec{I}_1 - \bar{F}_2\vec{\Phi}_2 + \sigma e^{-i\theta}H\vec{\Phi}_1 = 0, \tag{52}$$

where we recall that $F_1 \equiv z_1 + \gamma$, $\tilde{F}_1 \equiv (z_1 + \gamma')$, $F_2 \equiv \frac{1}{2}(F_1^2 + \delta)$, $H \equiv \rho' - \rho\tilde{F}_1 = -\rho(F_1 + \eta)$, and $\eta \equiv \gamma' - \gamma - \frac{\rho'}{\rho}$. We first analyze the case corresponding to $\rho = 0$. Solving this system we obtain the corresponding residue and DSII solution as

$$\vec{\Phi}_2 = \frac{1}{\Delta}\begin{pmatrix}\sigma F_1\rho' e^{-i\theta}\\ F_2\bar{F}_1\end{pmatrix}, \qquad q(x,y) = -\frac{2i\sigma\rho'}{\Delta}\ F_1 e^{-i\theta} \tag{53}$$

with $\Delta \equiv |F_2|^2 + |G|^2$. Finally Δ is now

$$\Delta \equiv |F_2|^2 + |H|^2 = \frac{1}{4}(\hat{y}^2 - \hat{x}^2 + \delta_R)^2 + (\hat{x}\hat{y} - \frac{\delta_I}{2} - t)^2 - \sigma|\rho'|^2. \tag{54}$$

In the moving frame with coordinates (49) we have

$$q(x,y) = -8i\sigma\rho'\frac{e^{2i(a\hat{x} - b\hat{y} + (b^2 - a^2)t + a\gamma_I + b\gamma_R)}(\hat{y} - i\hat{x})}{(\hat{y}^2 - \hat{x}^2 + \delta_R)^2 + 4(\hat{x}\hat{y} - \frac{\delta_I}{2} - t)^2 - 4\sigma|\rho'|^2}. \tag{55}$$

From a spectral point of view, and since ρ equals 0 this solution could be interpreted as arising from a one-dimensional homogeneous space of solutions and hence to correspond to the integers $m = N = d = 1$ and $Q = 2$. Note that, as has been commented, $\Phi_{11} = \frac{d}{dz}\log\Delta$ with $\Delta = (z\bar{z})^2, r^2 \to \infty$ (see (30), (33)). This solution for particular choice of the parameters was also obtained in [13] by use of Darboux formalism.

We consider next the general case corresponding to $\rho \neq 0$. Solving (52) we obtain the corresponding residue and DSII solution as

$$\vec{\Phi}_2 = \frac{1}{\Delta}\begin{pmatrix}\sigma(F_2\rho + F_1 H)e^{-i\theta}\\ F_2\bar{F}_1 + \sigma\bar{\rho}H\end{pmatrix}, \quad q(x,y) = -i\rho\sigma e^{-i\theta}\frac{(F_1^2 + 2\eta F_1 - \delta)}{\Delta}, \quad R = \ln\Delta, \tag{56}$$

and $\Delta \equiv |F_2|^2 - \sigma|H|^2$. In the moving frame with coordinates (49) we obtain

$$\Delta = \frac{1}{4}(\hat{y}^2 - \hat{x}^2 + \delta_R)^2 + \left(\hat{x}\hat{y} - \frac{\delta_I}{2} - t\right)^2 - \sigma|\rho|^2((\hat{x} - \eta_I)^2 + (\hat{y} + \eta_R)^2, \tag{57}$$

$$q(x,y,t) = -4i\sigma\rho e^{2i(a\hat{x}-b\hat{y}+(b^2-a^2)t+a\gamma_I+b\gamma_R)}$$

$$\cdot\frac{(\hat{y}^2-\hat{x}^2+2(\eta_R\hat{y}+\eta_I\hat{x})-\delta_R-2i(\hat{x}\hat{y}+\eta_R\hat{x}-\eta_I\hat{y}+\frac{\delta_I}{2}+t))}{(\hat{y}^2-\hat{x}^2+\delta_R)^2+4(\hat{x}\hat{y}-\frac{\delta_I}{2}-t)^2-4\sigma|\rho|^2((\hat{x}-\eta_I)^2+(\hat{y}+\eta_R)^2)}. \tag{58}$$

Spectrally this solution corresponds to the integers $m = N = 1$ and $d = Q = 2$.

Next we describe the main physical features of the above solution corresponding to $\sigma = -1$. The solution depends on 5 complex (or 10 real) parameters: $k_1, \gamma, \delta, \eta$, and ρ. It is nonsingular and at infinity it decays like $O(\frac{1}{r^2})$. The lump positions are asymptotically expected to be found at the locations at which the denominator has smallest order. Inspection of (54) shows that this occurs when the highest polynomials vanish at leading order. Hence we require $\hat{y}^2 - \hat{x}^2 = \hat{x}\hat{y} - t = 0$ and obtain for the lump positions

$$(\hat{x}_\pm, \hat{y}_\pm) \sim \pm\sqrt{|t|}(-1,1), \qquad t \to -\infty, \tag{59$-$}$$

and

$$(\hat{x}_\pm, \hat{y}_\pm) \sim \pm\sqrt{|t|}(1,1), \qquad t \to \infty, \tag{59$+$}$$

where $(\hat{x}_+, \hat{y}_+)$ and $(\hat{x}_-, \hat{y}_-)$ are the coordinates of the front and rear lumps. It follows that asymptotically the solution decomposes into two separate lumps, which, as seen in the Galilean frame (49), are moving with respect to the origin along the straight line $y = -x$ with velocities $\pm\frac{1}{2\sqrt{t}}(1,-1)$. Eventually they will collide and scatter off at an angle of $\frac{\pi}{2}$. This is unlike KPI, where the scattering angle takes on any possible value between 0 and 2π depending on the value of parameters [12]. The amplitude of both humps is maintained constant through the interaction process and given by $\frac{2}{|\rho|}$; this follows noting that at lump locations (59$\pm$) one has $\Delta = O(t)$, $F_1^2 + 2\eta F_1 = O(t)$, and hence their ratio is bounded with limit $\frac{2}{|\rho|}$. With respect to a frame at rest, the asymptotic motion of the lumps is more complicated. As $t \to \infty$ the trajectories follow a hyperbola given by $2(bx - ay)^2 + (a-b)(x-y) = 0$.

Lump scattering and motion can be thought of as the motion due to an attractive force $f(r) \approx \frac{1}{r^3}$ between the lumps. Define $x_{i,j} \equiv x_i - x_j, y_{i,j} \equiv y_i - y_j, i,j = +,-$, $\vec{r}_i \equiv (x_i, y_i), \vec{r}_{i,j} \equiv (x_{i,j}, y_{i,j}), r_{i,j} \equiv |\vec{r}_{i,j}|$. The above trajectories as $t \to \pm\infty$ of the humps solve the following system:

$$\frac{d^2\vec{r}_i}{dt^2} = \sum_{j,j\neq i}\frac{\vec{r}_{i,j}}{r_{i,j}}f(r_{i,j}).$$

We see that humps attract each other, but they do not form a bound state since the attractive interaction is not strong enough to bind them together. Figures 1, 2, and 3 show a plot of this configuration before, during, and after interaction.

We next study the regularity properties of the solution corresponding to $\sigma = -1$. The possible singularities of the solution appear at points at which the denominator vanishes. The denominator attains a minimum value $\mathcal{S}^2$ at $\hat{x} = \eta_I; \hat{y} = -\eta_R$ at a time $2t = -2\eta_I\eta_R - \delta_I \equiv 2\tilde{t}$, where $\mathcal{S} \equiv \eta_R^2 - \eta_I^2 + \delta_R$. At this point

$$|q|(\eta_I, -\eta_R, \tilde{t}) = \frac{4\rho}{|\mathcal{S}|}. \tag{60}$$

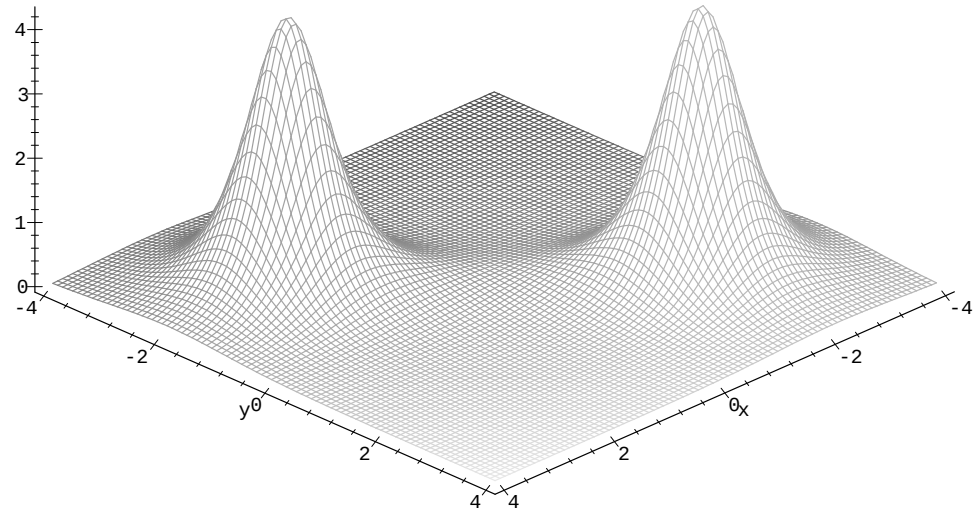


FIG. 1.

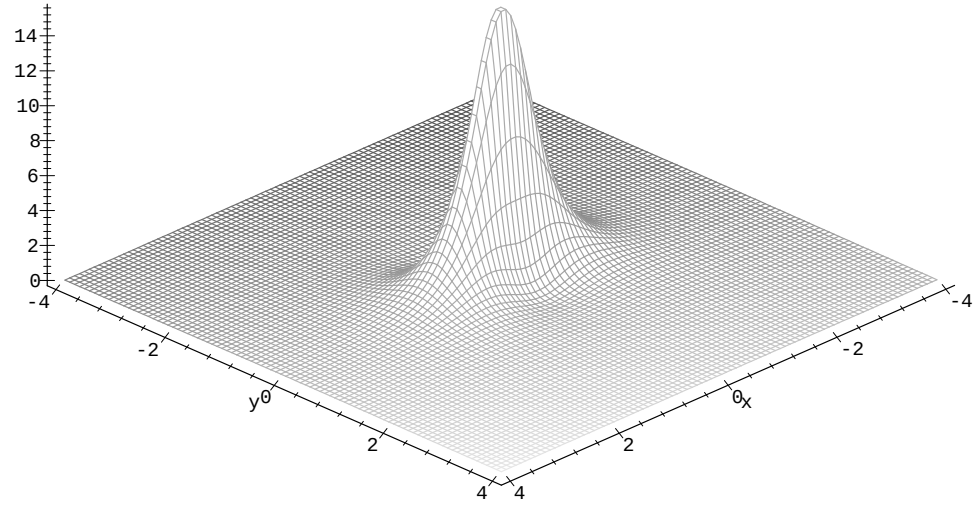


FIG. 2.

Thus the condition $\mathcal{S} \neq 0$ on the free parameters guarantees that the solution is never singular. However, one should not be misled into thinking that the amplitude grows unboundedly if the parameters satisfy $\mathcal{S} = 0$. The difficulty stems from letting first $\hat{x} = \eta_I$, $\hat{y} = -\eta_R$, $t = \tilde{t}$ and then $\mathcal{S} = 0$ (instead of doing these manipulations in the opposite order). Setting $t = \tilde{t}$, $\mathcal{S} = 0$ we find that (58) yields

$$|q|(\hat{x}, \hat{y}, \tilde{t}) = \frac{4\rho}{(\hat{x} + \eta_I)^2 + (\hat{y} - \eta_R)^2 + 4\rho^2}, \tag{60'}$$

and hence

$$|q|(\eta_I, -\eta_R, \tilde{t}) = \frac{\rho}{\eta_I^2 + \eta_R^2 + \rho^2} \leq |q|(-\eta_I, \eta_R, \tilde{t}) \equiv \frac{1}{\rho}.$$

This means that at time $t = \tilde{t}$ the two lumps have merged into one single rotationally

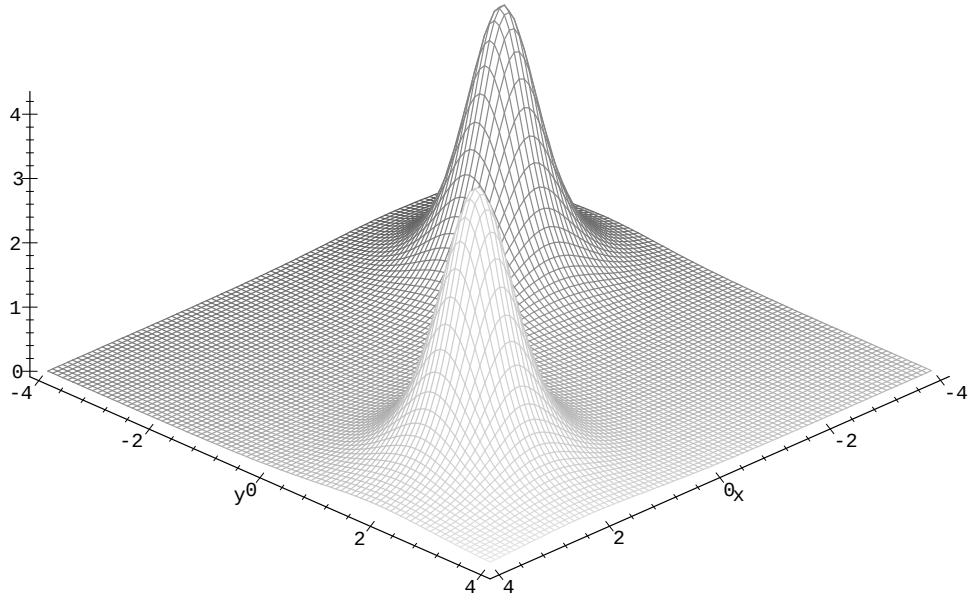


FIG. 3.

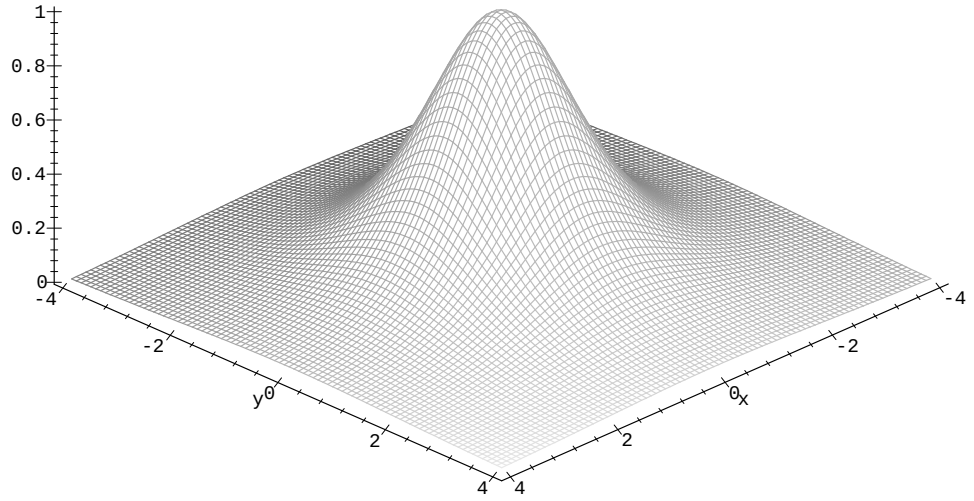


FIG. 4.

invariant structure and with an amplitude that is the sum of that of the individual lumps. This is interesting; it implies that all constants of motion of (58) depending only on the amplitude take the same value as those of (50) upon setting $\rho \to 2\rho$. In Figure 4 we show a plot of configuration during interaction $t = \tilde{t}$ displaying its rotational invariance.

The interaction process is similar to the solution (55). There exist several differences, however, between (55) and (58). We note that the DSII potential (55) decays as $\frac{1}{r^3}$ at infinity. Hence it has all the $L_p, p \geq 1$ norms finite. However, it depends only on 8 real parameters, and furthermore the lump's amplitude is not bounded but grows with time as $\sqrt{|t|}$ (recall that at the lump locations (55) one has $\Delta = O(1)$

and $F_1 = O(\sqrt{|t|}))$. Hence only for moderate periods of time can (55) be physically meaningful. On the other hand, (58) depends on 10 real parameters and decays at infinity (away from lump locations (55)) as $\frac{1}{r^2}$. The $L_p, p > 1$ norms are therefore finite—and hence there are all constants of the motion—but the L_1 norm is infinite. Unlike (55), the lump's amplitude of the lump (58) is bounded by $\frac{2}{|\rho|}$, and hence it can be physically meaningful for all times.

(iii) If $\mu(k)$ has the structure (45) and Q is any positive integer, then the Laurent coefficients satisfy (36.Q) with $\vec{\nu}_j^l = 0, l \geq 1, j = 1, 2$. Obtaining the corresponding DSII potential is a matter of linear algebra and is left for the reader.

Double poles. Next we briefly describe the simplest example of potentials corresponding to a double pole in the wave function. Assume that $\mu_1(k)$ has the representation

$$\vec{\mu}_1(k) = \frac{\vec{\Psi}_1}{(k-k_1)^2} + \frac{\vec{\Phi}_1}{k-k_1} + \vec{I}_1.$$

From Result 4 in such a case (39.1), (39.2) apply with $\vec{\nu}_1(k) = \vec{I}_1$. The resulting system can be solved to give the DSII potential (for convenience we take $\sigma = -1$)

$$q(x,y,t) = -2i\bar{\rho}e^{-i\theta}\frac{(\bar{F}_1^2 + 2\eta\bar{F}_1 - \delta)}{\Delta}$$

$$= -4i\bar{\rho}e^{i\theta}\frac{(\hat{y}^2 - \hat{x}^2 + 2(\eta_R\hat{y} + \eta_I\hat{x}) - \delta_R + 2i(\hat{x}\hat{y} + \eta_R\hat{x} - \eta_I\hat{y} + \frac{\delta_I}{2} - t))}{(\hat{y}^2 - \hat{x}^2 + \delta_R)^2 + 4(\hat{x}\hat{y} - \frac{\delta_I}{2} + t)^2 + 4|\rho|^2((\hat{x}-\eta_I)^2 + (\hat{y}+\eta_R)^2)}. \tag{61}$$

Spectrally this solution corresponds to the integers $N = 1$ and $m = d = Q = 2$. Note that this solution can be obtained from (58) upon letting $t \to -t$ and taking a complex conjugation. This is remarkable: it means that if $q(x,y)$ is a potential on the discrete spectrum of the DO corresponding to a wave function with simple poles and $Q = 2$, then $\bar{q}(x,y)$ is also related to the discrete spectrum of the DO corresponding to a wave function with double poles! Thus the transformation $q \to \tilde{q} \equiv \bar{q}(x,y,-t)$ preserves the index but not the pole structure. From a physical perspective this means that $q(x,y,t)$ and the physically related state $\tilde{q}(x,y,t) \equiv \bar{q}(x,y,-t)$, obtained by backwards evolution and conjugation of phase, have quite a different spectral classification in the case $Q = 2$. This stems from the following reason: the integers $N = 1$ and $m = d = Q = 2$ define an entire (equivalence) class of potentials, those $q_{k_1,\gamma,\delta,\eta,\rho}(x,y,t)$ members of the 10-parameter family (61). However, $\tilde{q}$ is not a member of the latter family, unlike what happens for $Q = 1$, where $\tilde{q}_{k_1,\gamma,\rho}(x,y,t) = q_{-k_1,\gamma,-\rho}(x,y,t)$. It follows that for $Q = 2$, $\tilde{q}$ cannot correspond to the same spectral numbers that q does.

Potentials corresponding to superposition of simple poles. More complicated solutions are found by taking a simple linear combination of N poles located at $k_l \equiv a_l + ib_l, l = 1, \ldots, N$. Assume that

$$\vec{\mu}_1(k) = \vec{I}_1 + \sum_{l=1}^{N}\frac{\vec{\Phi}^{l)}}{(k-k_l)}, \qquad \vec{\mu}_2(k) = \vec{I}_2 + \sum_{l=1}^{N}\frac{\tau\vec{\Phi}^{l)}}{(k-\bar{k}_l)}. \tag{62}$$

In this case one has that at every pole k_j

$$\vec{\nu}_1(k_j) = \vec{I}_1 + \sum_{l\neq j}^{N} \frac{\vec{\Phi}^{l)}}{a_{jl}}, \qquad \vec{\nu}_1^1(k_j) = -\sum_{l\neq j}^{N} \frac{\vec{\Phi}^{l)}}{a_{jl}^2}, \tag{63.1}$$

$$\vec{\nu}_2(\bar{k}_j) = \vec{I}_2 + \sum_{l\neq j}^{N} \frac{\tau\vec{\Phi}^{l)}}{\bar{a}_{jl}}, \qquad \vec{\nu}_2^1(\bar{k}_j) = -\sum_{l\neq j}^{N} \frac{\tau\vec{\Phi}^{l)}}{\bar{a}_{jl}^2}, \tag{63.2}$$

where $a_{jl} \equiv k_j - k_l$. We obtain the following:

(i) Assume that $Q^{j)} = 1, j = 1, \dots, N$. Then the solution satisfies

$$|q|^2 = -\sigma(\partial_{xx} + \partial_{yy})R, \qquad R(x,y) = \log\Delta. \tag{64}$$

Here $\Delta \equiv \det B$ is real, B is the $2N \times 2N$ matrix with the block decomposition

$$B = \begin{pmatrix} m & n \\ \sigma\bar{n} & \bar{m} \end{pmatrix}, \tag{65}$$

and we have defined

$$m_{jl} = F_1^{j)}\delta_{jl} - (1-\delta_{jl})\frac{e^{-a_{jl}z}}{a_{jl}}, \qquad n_{jl} = \bar{\rho}^{j)}\delta_{jl}, \qquad j,l = 1,2,\dots,N, \tag{66}$$

$$F_1^{j)} = z_1 + \gamma^{j)}, \quad z_1 \equiv z \equiv y - ix, \quad \theta^{j)} \equiv \theta_{21}^{j)} = 2(b^{j)}y - a^{j)}x + ({b^{j)}}^2 - {a^{j)}}^2)t).$$

(Note that we use superscripts to label functions corresponding to different poles.)

The solution is a rational function with coefficients modulated by the difference of phases $\theta^{j)} - \theta^{k)}, j,k = 1,\dots,N$; for long times it is a superposition of N lumps like (50), each of them traveling with a speed that for the $^{j)}$th lump is given by $(-2a^{j)}, -2b^{j)})$. These lumps asymptotically do not suffer interaction effects among themselves. Finally

$$\|q\|_2^2 = 4\pi N. \tag{67}$$

(ii) All the $Q^{j)} = 2$. In this case let

$$B = \begin{pmatrix} m & n \\ \sigma\bar{n} & \bar{m} \end{pmatrix}, \qquad \Delta = \det B, \tag{68}$$

where we define the $N \times N$ matrices m and n by

$$m_{jl} \equiv F_2^{j)}\delta_{jl} - \left[\frac{F^{j)}}{a_{jl}} + \frac{1}{a_{jl}^2}\right] e^{-a_{jl}z}(1-\delta_{jl}), \quad n_{jl} \equiv -\bar{H}^{j)}\delta_{jl} - \bar{\rho}^{j)}\frac{e^{-\bar{a}_{jl}\bar{z}}}{\bar{a}_{jl}}(1-\delta_{jl}). \tag{69}$$

Then

$$|q|^2 = -\sigma(\partial_y^2 + \partial_x^2)\log\Delta, \qquad R(x,y) = \log\Delta, \tag{70}$$

where Δ is real and

$$\|q\|_2^2 = 8\pi N. \tag{71}$$

The solution is a rational function with coefficients depending on the difference of phases $e^{i(\theta^{j)}-\theta^{k)})}, j,k=1,\dots,N$. For long time, the solution is composed of a sum of $2N$ lumps like (58); the trajectory for the $^{j)}$th lump is given by (59) with the relevant parameters. Only pairs of lumps corresponding to the same pole interact, and in a similar way to that described above with scattering in an angle of $\frac{\pi}{2}$. Otherwise, lumps corresponding to different poles do not interact.

Proof. We first prove that Δ is real (note that the proof is valid for both case (i) and case (ii)). We note the following:

$$\Delta = \det\begin{pmatrix} m & n \\ \sigma\bar{n} & \bar{m}\end{pmatrix} = (1)^N \det\begin{pmatrix} \sigma\bar{n} & \bar{m} \\ m & n\end{pmatrix} = \det\begin{pmatrix} \bar{m} & \sigma\bar{n} \\ n & m\end{pmatrix}$$
$$= \det\begin{pmatrix} \bar{m} & \bar{n} \\ \sigma n & m\end{pmatrix} = \overline{\Delta}.$$

We next prove formula (64). If all the $Q^{j)}=1$, then equations (36.1) apply and for $j=1,\dots,N$ we obtain the system

$$F_1^{j)}\Phi^{j)} - \sum_{l\neq j}^{N} \frac{\vec{\Phi}^{l)}}{a_{jl}} + \bar{\rho}^{j)} e^{i\theta^{j)}} \tau\vec{\Phi}^{j)} = \vec{I}_1,$$

$$\bar{F}_1^{j)}\tau\Phi^{j)} - \sum_{l\neq j}^{N} \frac{\tau\vec{\Phi}^{l)}}{\bar{a}_{jl}} + \sigma\rho^{j)} e^{-i\theta^{j)}} \vec{\Phi}^{j)} = \vec{I}_2. \tag{72}$$

For convenience we introduce the column vector $\vec{p}$ with entries $p^{j)} \equiv e^{-k^{j)}z}$ and

$$\vec{\pi}^{j)} = \vec{\Phi}^{j)} p^{j)}, \quad \zeta^{j)} \equiv \vec{\pi}_1^{j)}, \quad \omega^{j)} \equiv \sigma\bar{\pi}_2^{j)},$$

in terms of which (72) reads

$$\sum_{l\neq j}^{N} m_{jl}\zeta^{l)} + \bar{\rho}^{j)}\omega^{j)} = e^{-k^{j)}z}, \qquad \sum_{l\neq j}^{N} \bar{m}_{jl}\omega^{l)} + \sigma\rho^{j)}\zeta^{j)} = 0.$$

If B is the $2N\times 2N$ matrix defined in (65) and for $j=1,\dots,n$, $B^{j)}$ is the $2N\times 2N$ matrix obtained substituting the $^{j)}$th column of B by the column vector $\left(\begin{smallmatrix}\vec{p}\\ \vec{0}\end{smallmatrix}\right)$, the above system can be written as

$$B\begin{pmatrix}\zeta^{j)}\\ \omega^{j)}\end{pmatrix} = \begin{pmatrix}\vec{p}\\ \vec{0}\end{pmatrix}.$$

Therefore

$$\zeta^{j)} = \frac{B^{j)}}{B}, \qquad \sum_{j=1}^{N}\Phi_{j1} = \frac{1}{B}\sum_{j=1}^{N}\frac{B^{j)}}{p^{j)}} = \frac{\partial}{\partial z}\log\Delta,$$

where the last equality is obtained using the expression for the derivative of a determinant along with the fact

$$\frac{\partial}{\partial z}m_{jl}(z) = \frac{p^{l)}}{p^{j)}}, \qquad \frac{\partial}{\partial z}\bar{m}_{jl}(z) = \frac{\partial}{\partial z}\sigma\bar{n}_{jl}(z) = \frac{\partial}{\partial z}n_{jl}(z) = 0$$

and on account of (25.2)

$$|q|^2 = -4\sigma \frac{\partial}{\partial \bar{z}} \frac{\partial}{\partial z} \log \Delta.$$

This solution was obtained in [4]. Unlike what (64)–(66) suggests (and as was claimed in [4]), for $N > 1$, Δ does not increase exponentially; to see this note that $\Delta = \det B'$, where B' is the $2N \times 2N$ matrix with the block decomposition

$$B' = \begin{pmatrix} m' & n' \\ \sigma \bar{n}' & \bar{m}' \end{pmatrix},$$

and we have defined

$$m'_{jl} = F_1^{j)} \delta_{jl} - (1 - \delta_{jl}) \frac{1}{a_{jl}}, \quad n'_{jl} = \bar{\rho}^{j)} e^{i\theta^{j)}} \delta_{jl}, \quad j, l = 1, 2, \dots, N.$$

From this expression it can be proven that Δ is a polynomial with coefficients depending on $e^{i(\theta^{j)} - \theta^{k)})}, j, k = 1, \dots, N$. It also follows that for long values of either z or t, $\Delta = \prod_{j=1}^{N} F_1^{j)} \bar{F}_1^{j)} + O(z\bar{z})^{N-2}$; i.e., the relevant determinant factorizes as a product of determinants corresponding to one lumps, with relevant parameters, and hence that the solution is a superposition of N lumps like (50). In addition, (32) and $\Delta = (z\bar{z})^N + O(z\bar{z})^{N-2}$ imply that $\|q\|_2^2 = 4\pi N$, i.e., (67).

The case corresponding to all the $Q^{j)} = 2$ can be proven along similar lines using equations (36).

Summary of potentials. Below we summarize the above discussion on the spectral classification of potentials on the discrete spectrum and DSII solutions. Recall that the integers N, m, Q, d are defined at the beginning of section 3. Finally s represents the number of fundamental lumps the solution is composed of.

Spectral classification.

Solution	N	m	Q	d	s
(50)	1	1	1	2	1
(55)	1	1	2	1	2
(58)	1	1	2	2	2
(61)	1	2	2	2	2
(66)	N	1	1	2	N
(69)	N	1	2	2	$2N$

Physical properties. The following table gives a summary of the physical properties of the main localized DSII solutions (we take $\sigma = -1$). Recall that $\|q\|_2^2 = 4\pi s$ is the (square of the) L_2 norm, which physically represents the mass of the wave.

Solution	s	Smooth	Scattering	Angle	Amplitude	Decay	Parameters	Mass
(50)	1	Yes			$\frac{2}{\rho}$	$1/r^2$	5	4π
(55)	2	Yes	Yes	$\frac{\pi}{2}$	$\sqrt{t} \to \infty$	$1/r^3$	8	8π
(58)	2	Yes	Yes	$\frac{\pi}{2}$	$\frac{2}{\rho}$	$1/r^2$	10	8π
(61)	2	Yes	Yes	$\frac{\pi}{2}$	$\frac{2}{\rho}$	$1/r^2$	10	8π
(66)	N	Yes	No	0	$\frac{2}{\rho}$	$1/r^2$	$5N$	$4\pi N$
(69)	$2N$	Yes	Yes (by pairs)	$\frac{\pi}{2}$	$\frac{2}{\rho}$	$1/r^2$	$10N$	$8\pi N$

Remarks. We have presented the simplest examples of potentials associated with the discrete spectrum and localized solutions to DSII. While we shall not elaborate any further on this, we remark that one could also consider a mixture of poles of different types. Also, following the methods described one can in principle derive equations corresponding to higher order charges and/or poles.

Potentials corresponding to general spectrum. Next we consider potentials corresponding to the general case when both continuous and discrete spectrums are present. For simplicity we assume that the discrete spectrum corresponds to having just one simple pole; more general cases can be handled via the theory exposed before and linear algebra. One has the following:

Assume that there exists a solution to (7) such that its columns $\vec{\mu}_1(k), \vec{\mu}_2(k)$ have simple poles $k_1 \equiv a + ib, \bar{k}_1$ and residues $\vec{\Phi}, \tau\vec{\Phi}$, respectively,

$$\vec{\mu}_1(k) = \vec{\mu}_{1\mathrm{reg}}(k) + \frac{\vec{\Phi}}{(k-k_1)}, \qquad \vec{\mu}_2(k) = \tau\vec{\mu}_{1\mathrm{reg}}(k) + \frac{\tau\vec{\Phi}}{(k-\bar{k}_1)}, \tag{73}$$

where $\vec{\mu}_{1\mathrm{reg}}(k) \neq \vec{I}_1$ and $\vec{\mu}_{1\mathrm{reg}}(k) \to \vec{I}_1$ as $k \to \infty$. Then, in terms of the definitions given above, $\vec{\mu}_1(k)$ solves the linear equation of the inverse problem

$$\vec{\mu}_1(k) = \vec{I}_1 + \frac{1}{2\pi i}\int_C \frac{e^{i\theta_{12}}T_{12}(z)}{z-k}\tau\vec{\mu}_1(\bar{z})dz\wedge d\bar{z} + \frac{\vec{\Phi}}{(k-k_1)}, \tag{74}$$

and also, if $Q = 1$,

$$\vec{I}_1 + \frac{1}{2\pi i}\int_C \frac{e^{i\theta_{12}}T_{12}(z)}{z-k_1}\tau\vec{\mu}_1(\bar{z})dz\wedge d\bar{z} = F_1\vec{\Phi} + \bar{\rho}e^{i\theta}\tau\vec{\Phi} \tag{75.1}$$

or, if $Q = 2$,

$$\frac{\partial}{\partial k_1}\int_C \frac{e^{i\theta_{12}}T_{12}(z)}{z-k_1}\tau\vec{\mu}_1(\bar{z})dz\wedge d\bar{z} = F_1\left[2\pi i\vec{I}_1 + \int_C \frac{e^{i\theta_{12}}T_{12}(z)}{z-k_1}\tau\vec{\mu}_1(\bar{z})dz\wedge d\bar{z}\right]$$

$$+\bar{\rho}e^{i\theta}\left[2\pi i\vec{I}_2 + \sigma\int_C \frac{e^{-i\theta_{12}}\bar{T}_{12}(z)}{\bar{z}-\bar{k}_1}\vec{\mu}_1(z)dz\wedge d\bar{z}\right] - 2\pi i F_2\vec{\Phi} + 2\pi i e^{i\theta}\bar{H}\tau\vec{\Phi}. \tag{75.2.}$$

Finally the potential is obtained from (25) with

$$\xi_{12} = \frac{\sigma}{2\pi i}\int_C e^{-i\theta_{12}}\bar{T}_{12}(z)\mu_{11}(\bar{z})dz\wedge d\bar{z} + \sigma\bar{\Phi}_2. \tag{76}$$

Proof. If $\mu(k)$ has the structure (73), it follows that there exists a solution $\vec{\Phi}$ to the homogeneous equation $\mathcal{G}_1(k_1)\vec{\Phi} = \vec{0}$ at $k = k_1$. In this case one finds that (10) is to be modified as follows:

$$\frac{\partial\mu}{\partial\bar{k}} = \sum_{j,l=1}^{n}\mu\left(k_R + i\frac{J_j}{J_l}k_I\right)\Omega^{lj}(k) + A\Phi\delta(k-k_j), \tag{10$'$}$$

where A is an arbitrary constant. Direct derivation of (73) and use of the relationship $\frac{\partial}{\partial\bar{k}}\frac{1}{(k-k_j)} = \pi\delta(k-k_j)$ shows that $A = \pi$. Equation (74) follows substituting (10′) in (11). To obtain (75) use (36)—which is valid for general wave functions that have

only simple poles—the symmetry relationship (18), (19), and also the fact that (74) implies that

$$\vec{\nu}_1(k_1) = \vec{I}_1 + \frac{1}{2\pi i}\int_C \frac{e^{i\theta_{12}}T_{12}(z)}{z-k_1}\tau\vec{\mu}_1(\bar{z})dz\wedge d\bar{z},$$

$$\vec{\nu}_1^1 = \frac{1}{2\pi i}\frac{\partial}{\partial k_1}\int_C \frac{e^{i\theta_{12}}T_{12}(z)}{z-k_1}\tau\vec{\mu}_1(\bar{z})dz\wedge d\bar{z}.$$

Equation (76) also follows from (73).

Remark. We leave it to the reader to derive the linear equation of the inverse problem that $\vec{\mu}_1(k)$ satisfies when it has the structure (73) and Q is any positive integer. One only needs to determine $\vec{\nu}_j^l, l \geq 1, j = 1, 2$, and note that in this case the Laurent coefficients satisfy (36.Q).

5. The general case. Most of the former theory can be generalized to the general case corresponding to $J = \text{ Diag}(J_1, \dots, J_n)$, $J_1 \neq J_2 \neq \cdots \neq J_n$, and $A(x, y)$ an off-diagonal matrix. We mention the relevant generalizations.

Consequence of Result 0. For any fixed column j, let $k_j \equiv a_j + ib_j$ be an eigenvalue for that column: i.e., there exists a solution $\vec{\omega}_j$ to the jth homogeneous equation at the point k_j and let $\mathcal{L}_j$ be the set of indices

$$\mathcal{L}_j = \left\{ l| \text{ exists } \vec{\omega}_l \text{ that solves the } l\text{th homogeneous equation at } a_j + i\frac{J_j b_j}{J_l} \equiv k_{lj} \right\}.$$

Let $\pi_{lj} \equiv e^{-i\theta_{jl}(k_{lj})}\vec{\omega}_l$. Note that $j \in \mathcal{L}_j \subset \{1, \dots, n\}$ and hence in particular (if $l = j$) $\pi_{jj} = \vec{\omega}_j$. Then the null space of the jth homogeneous equation at $k = k_j$ contains the span of the functions π_{lj} when l ranges in $\mathcal{L}$:

$$\{\pi_{lj}\}_{l\in\mathcal{L}} \subset \text{ Ker } \mathcal{G}_j(k_j) =; \quad \text{Dim Ker } \mathcal{G}_j(k_j) \equiv |\mathcal{L}| \geq 1.$$

Result 3 is generalized as follows.

Let $\vec{\mu}_j(k)$ have the Laurent expansion (13) around any simple pole k_1. Then the following hold:

(i) The index matrix must be diagonal: $(Q)_{lj} = \text{ Diag } (Q_1, \dots, Q_n)$.

(ii) If $Q_j = 1$, then the Laurent coefficients satisfy

$$\vec{\nu}_j = F_{j;1}\vec{\Phi}_j + \sum_{l\in\mathcal{L}} \rho_{jl}e^{-i\theta_{jl}(k_l)}\vec{\Phi}_l, \tag{77.1}$$

where $F_{j;1} \equiv (z_j + \gamma_j), F_{j;2} \equiv \frac{1}{2}(F_{j;1}^2 + \delta_j)$, and $\gamma, \delta, \rho_{jl}, \dots$ are arbitrary constants.

(iii) If $Q_j = 2$, then the Laurent coefficients satisfy

$$\vec{\nu}_j^1 = F_{j;1}\vec{\nu}_j - F_{j;2}\vec{\Phi}_j + \sum_{l\in\mathcal{L}} e^{-i\theta_{jl}(k_l)}\left(\rho_{jl}\vec{\nu}_l + (\rho'_{jl} - \rho_{jl}\tilde{F}_{l;1})\vec{\Phi}_l\right), \tag{77.2}$$

where $\tilde{F}_{j;1} \equiv (z_j + g'_j)$ and g'_j, ρ'_{jl} are new constants.

Temporal evolution. In the general case the relevant evolution equation is the compatibility of (1) and

$$\left[\partial_t - iJ_l\partial_{yy} + 2ik_jJ_l\partial_y + ik_j^2(J_j - J_l)\right]\mu_{lj} + \sum_r A_{lr}\partial_y\mu_{rj} - A_{1lr}\mu_{rj} + k_jA_{lr}\mu_{rj} = 0,$$
$$l, j = 1, \dots, n,$$

where A_{1lr} is a certain matrix. One finds that the discrete scattering data corresponding to an eigenvalue $k_j \equiv a_j + ib_j$ evolves as the following.

Simple poles.

$$\partial_t \gamma_j = -2iJ_j k_j, \quad \partial_t \delta_j = 2iJ_j, \quad \partial_t \rho_{lj} = iJ_{lj}\left(a_l^2 + i\frac{J_l}{J_j}b_l^2\right)\rho_{lj},$$

$J_{lj} \equiv J_l - J_j$ and hence that $\gamma_j(t) = \gamma_j(0) - 2iJ_j k_j t$, $\delta_j(t) = \delta_j(0) + 2iJ_j t$,

$$\rho_{lj}(t) = \rho_{lj}(0)\exp iJ_{lj}\left(a_l^2 + i\frac{J_l}{J_j}b_l^2\right). \tag{78}$$

Double poles.

$$\gamma_j(t) = \gamma_j(0) - 2iJ_j k_j t, \qquad \delta_j(t) = \delta_j(0) - 2iJ_j t, \tag{79}$$

$$\rho_{lj}(t) = \rho_{lj}(0)\exp iJ_{lj}\left(a_l^2 + i\frac{J_l}{J_j}b_l^2\right). \tag{80}$$

Appendix. Here we prove that for DSII, no poles in $\bar{k}$ may exist.

Assume the eigenfunction has canonical normalization and the following expansion as $|k| \to \infty$:

$$\mu(k) = I + \frac{\Phi}{k} + \frac{\varphi}{\bar{k}} + o(1/k). \tag{A.1}$$

Inserting this expansion into (3) one obtains that (we consider $n = 2$, i.e., DSII)

$$\bullet O(k): \quad (J_j - J_l)\delta_{lj} = 0,$$

which is identically satisfied.

$$\bullet O(1): \quad A_{lj} = i(J_j - J_l)\Phi_{lj}, \tag{A.2}$$

which implies

$$A_{ii} = 0, \tag{A.3}$$

$$\bullet O(k/\bar{k}): \quad (J_j - J_l)\varphi_{lj} = 0. \tag{A.4}$$

It follows that

$$\varphi_{ij} = 0, \qquad i \neq j, \tag{A.5}$$

$$\bullet O(1/\bar{k}): \quad (\partial_x + iJ_l\partial_y)\,\varphi_{lj} = \sum_r A_{lr}\varphi_{rj} = A_{lj}\varphi_{jj}.$$

With $l \neq j$ we have that the left-hand side is zero: $(\partial_x + iJ_l\partial_y)\,\varphi_{lj} = 0$ (using (A.5)); since for off-diagonal elements $A_{lj} \neq 0, l \neq j$, we conclude that

$$\varphi_{jj} = 0 \qquad \text{and} \qquad \varphi_{lj} = 0 \text{ for all } l, j. \tag{A.6}$$

Assume a wave function with canonical normalization and pure simple poles at $k_1, \bar{k}_1$ in both variables $k, \bar{k}$ exists, i.e., that

$$\vec{\mu}_1(k) = \vec{I}_1 + \frac{\vec{\Phi}_1}{(k-k_1)} + \frac{\vec{\varphi}_1}{(\bar{k}-\bar{k}_1)} \tag{A.7}$$

and (recalling that we consider DSII)

$$\vec{\mu}_2(k) = \vec{I}_2 + \frac{\tau\vec{\Phi}_1}{(k-\bar{k}_1)} + \frac{\tau\vec{\varphi}_1}{(\bar{k}-k_1)}. \tag{A.8}$$

It follows that this eigenfunction has, as $|k| \to \infty$, the expansion (A.1) with

$$\varphi = \begin{pmatrix} \vec{\varphi}_1 & \tau\vec{\varphi}_1 \end{pmatrix} = \begin{pmatrix} \varphi_{11} & \sigma\bar{\varphi}_{21} \\ \varphi_{21} & \bar{\varphi}_{11} \end{pmatrix}.$$

It follows that $\vec{\varphi}_1 = \vec{0}$.

Therefore, no "pure" poles eigenfunctions of the type (A.7) exist; we expect that neither will do more complicated pole states.

This argument must still apply when a continuous spectrum is added; indeed, in the scattering theory the discrete spectrum separates from the continuous spectrum—so it seems unnecessary to consider the latter. This is substantiated using a formal expansion of the continuous spectrum in formula (74), which does not contain terms in $1/\bar{k}$.

REFERENCES

[1] A. Davey and K. Stewartson, *On three-dimensional packets of surface waves*, Proc. Roy. Soc. London Ser. A, 338 (1974), pp. 101–110.

[2] M. J. Ablowitz and H. Segur, *On the evolution of packets of water waves*, J. Fluid Mech., 92 (1979), pp. 691–715.

[3] K. Nishinari, K. Abe, and J. Satsuma, J. Phys. Soc. Japan, 62 (1993), p. 2021.

[4] V. A. Arkadiev, A. K. Progrebkov, and M. C. Polivanov, *Inverse scattering transform method and soliton solutions for Davey–Stewartson* II *equation*, Phys. D, 36 (1989), pp. 189–197.

[5] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, London Math. Soc. Lecture Note Ser. 149, Cambridge University Press, Cambridge, UK, 1989.

[6] M. J. Ablowitz and A. S. Fokas, *Comments on the inverse scattering transform and related nonlinear evolution equations*, in Nonlinear Phenomena (Oaxtepec, 1982), Lecture Notes in Phys. 189, Springer-Verlag, Berlin, 1983, pp. 3–24. See also [7].

[7] A. S. Fokas and M. J. Ablowitz, *The inverse scattering transform for multidimensional* $(2+1)$ *problems*, in Nonlinear Phenomena (Oaxtepec, 1982), Lecture Notes in Phys. 189, Springer-Verlag, Berlin, 1983, pp. 137–183.

[8] A. Fokas and M. J. Ablowitz, *On the inverse scattering transform of multidimensional nonlinear equations related to first-order systems in the plane*, J. Math. Phys., 25 (1984), pp. 2494–2505.

[9] J. Villarroel and M. J. Ablowitz, *On the Hamiltonian formalism for the Davey–Stewartson system*, Inverse Problems, 7 (1991), pp. 451–460.

[10] M. J. Ablowitz and J. Villarroel, *On the complete integrability of certain nonlinear evolution equations in one and two spatial dimensions*, in Chaos & Order (Canberra, 1990), World Scientific, Teaneck, NJ, 1991, pp. 1–13.

[11] M. J. Ablowitz and J. Villarroel, *Solutions to the time dependent Schrödinger and the Kadomtsev–Petviashvili equations*, Phys. Rev. Lett., 78 (1997), pp. 570–573.

[12] J. Villarroel and M. J. Ablowitz, *On the discrete spectrum of the nonstationary Schrödinger equation and multipole lumps of the Kadomtsev–Petviashvili* I *equation*, Comm. Math. Phys., 207 (1999), pp. 1–42.

[13] M. Mañas and P. Santini, *Solutions of the Davey–Stewartson* II *equation with arbitrary rational localization and nontrivial interaction*, Phys. Lett. A, 227 (1997), pp. 325–334.
[14] R. S. Ward, *Nontrivial scattering of localized solitons in a* $(2 + 1)$*-dimensional integrable system*, Phys. Lett. A, 208 (1995), pp. 203–208.
[15] K. A. Gorshkov, D. E. Pelinovsky, and Yu. A. Stepanyants, *Normal and anomalous scattering, formation and decay of bound states of two-dimensional solitons described by the Kadomtsev–Petviashvili equation*, JETP, 77 (1993), pp. 237–245.
[16] M. J. Ablowitz, S. Chakravarty, A. D. Trubatch, and J. Villarroel, *A novel class of solutions of the non-stationary Schrödinger and the Kadomtsev–Petviashvili* I *equations*, Phys. Lett. A, 267 (2000), pp. 132–146.
[17] D. Pelinovsky, *Rational solutions of the KP hierarchy and the dynamics of their poles.* II. *Construction of the degenerate polynomial solutions*, J. Math. Phys., 39 (1998), pp. 5377–5395.
[18] Q. P. Liu and M. Mañas, *Vectorial Darboux transformations for the Kadomtsev–Petviashvili hierarchy*, J. Nonlinear Sci., 9 (1999), pp. 213–232.
[19] A. S. Fokas and L.-Y. Sung, *On the solvability of the N-wave, Davey–Stewartson and Kadomtsev–Petviashvili equations*, Inverse Problems, 8 (1992), pp. 673–708.
[20] L.-Y. Sung, *An inverse scattering transform for the Davey–Stewartson* II *equations.* I, J. Math. Anal. Appl., 183 (1994), pp. 121–154.
[21] L.-Y. Sung, *An inverse scattering transform for the Davey–Stewartson* II *equations.* II, J. Math. Anal. Appl., 183 (1994), pp. 289–325.
[22] L.-Y. Sung, *An inverse scattering transform for the Davey–Stewartson* II *equations.* III, J. Math. Anal. Appl., 183 (1994), pp. 477–494.
[23] D. E. Pelinovsky and C. Sulem, *Spectral decomposition for the Dirac system associated to the DSII equation*, Inverse Problems, 16 (2000), pp. 59–74.
[24] M. J. Ablowitz and A. S. Fokas, *Complex Variables: Introduction and Applications*, Cambridge University Press, Cambridge, UK, 1997.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1279–1299

EXISTENCE OF WEAK SOLUTIONS TO SOME VORTEX DENSITY MODELS*

QIANG DU† AND PING ZHANG‡

Abstract. We study the weak solutions to equations arising in the modeling of vortex motions in superfluids such as superconductors. The global existence of measure-valued solutions is established with a bounded Radon measure as initial data. Moreover, we get a local space-time L^q estimate for the continuous part of the solution, and we prove the global existence of a distributional weak solution for a particular case. We also consider a modification to the model in order to physically account for the different signs of vortices, and we present, in one space dimension, the global existence of weak solutions with the initial data in BV for the modified model.

Key words. quantized vortices, vortex density, hydrodynamics, vortex sheets, weak convergence, measure-valued solutions

AMS subject classifications. 35Q55, 35B, 35K

PII. S0036141002408009

1. Introduction.

In this paper, we study the concentration phenomenon of the approximate solution sequences to the equations

$$\begin{cases} \partial_t \rho + \operatorname{div}(u\rho) = 0, \quad (t,x) \in (0,\infty) \times \mathbb{R}^2, \\ u = M\nabla \triangle^{-1} \rho, \\ \rho|_{t=0} = \rho_0, \end{cases} \tag{1.1}$$

with ρ_0 being a bounded Radon measure and M being a constant orthogonal matrix of the form

$$M(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Our investigation yields the global existence of a measure-valued solution to (1.1) and the classical weak solution to (1.1) if ρ_0 is a bounded positive (resp., negative) Radon measure when $\cos\theta > 0$ (resp., $\cos\theta < 0$). For the case $\cos\theta = 1$, our results here extend those available in the literature (see, for instance, [21]). In the more general case, our study is related to the mathematical study of incompressible fluids as well as the vortex state in superfluids.

Indeed, when $\cos\theta = 0$, (1.1) is the classical two-dimensional incompressible Euler equations, which can be rewritten in the velocity formulation

$$\begin{cases} \partial_t u + \operatorname{div}(u \otimes u) = -\nabla P, \quad (t,x) \in (0,\infty) \times \mathbb{R}^2, \\ \operatorname{div} u = 0, \\ u|_{t=0} = u_0. \end{cases} \tag{1.2}$$

*Received by the editors May 21, 2002; accepted for publication (in revised form) November 18, 2002; published electronically May 12, 2003. This work was supported in part by the Chinese NSF, innovation grants from the Chinese Academy of Sciences, the state key basic research project G199903280, and the U.S. NSF grant DMS-0196522. This work was completed during Ping Zhang's visit to Penn State University.

http://www.siam.org/journals/sima/34-6/40800.html

†Department of Mathematics, Penn State University, University Park, PA 16802, and Lab for Scientific and Engineering Computing, CAS, China (qdu@math.psu.edu).

‡Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing, 100080, China (zp@mail.math.ac.cn).

The initial value problem to (1.2) with $u_0 = (u_0^1, u_0^2) \in L^2_{loc}(\mathbb{R}^2)$ and $\omega_0 = \partial_2 u_0^1 - \partial_1 u_0^2 \in \mathcal{M}(\mathbb{R}^2)$ is an outstanding open problem, known as the vortex sheets problem, in incompressible fluid mechanics (see [25]). In 1991, Delort [7] proved the global existence of weak solutions to this problem when ω_0 is a bounded Radon measure without negative singular part. Later, Majda [26] obtained the same result by the vanishing viscosity limit to the two-dimensional incompressible Navier–Stokes equations. When ω_0 is a general Radon measure, only a measure-valued solution seems possible for this problem; see [8, 9, 27, 28] for more details.

In recent years, studies of the stability, dynamics, and interactions of the vortices in both classical fluids and superfluids have received a lot of attention. For example, in the mezoscale Ginzburg–Landau models of superconductors [14], the individual vortices are resolved and their interactions and dynamics are studied in great detail. On a macroscopic level, when the number of vortices becomes exceedingly large, it is advantageous to model the vortex state using a vortex density function [4, 16]. When $\cos\theta = 1$, (1.1) has been obtained as the hydrodynamic limit of Ginzburg–Landau vortices governing by gradient dynamics. A formal derivation was given in [16], and a rigorous justification was given in [21]. Studies in this direction also include [2, 4, 15] on model derivation, [19, 29, 30] on mathematical analysis, and [5, 12, 13, 18] on numerical simulations (see [3] for additional references). The general case of $\theta \neq 0$ corresponds to a complex time relaxation in the gradient dynamics. In [21], the global existence of weak solutions to (1.1) with $\cos(\theta) = 1$ and ρ_0 being a positive bounded Radon measure was established. In the case of ρ_0 taking on different signs, the notion of weak solutions to (1.1) with $\cos(\theta) \neq 0$ requires further study, as additional difficulties do arise.

A similar but somewhat modified version of (1.1) was studied in [2, 4] when the vortices are of different signs. Taking the London approximation to the induced magnetic field into account, a system of equations similar to (1.1) with $\cos(\theta) = 1$ was derived in [4]:

$$\begin{cases} \partial_t \rho + \operatorname{div}(u|\rho|) = 0, \quad (t,x) \in (0,\infty) \times \mathbb{R}^2, \\ u = \nabla(\lambda^2 \triangle - I)^{-1} \rho, \\ \rho|_{t=0} = \rho_0. \end{cases} \tag{1.3}$$

Here, λ denotes the penetration depth. The density function ρ is allowed to change sign in order to represent vortices of different signs. The general case of $\theta \neq 0$ can also be easily derived when the time relaxation parameter becomes complex valued [17]. In [11], such an approach was taken to account for the Hall effect. A vector-valued version of (1.3) was also available [4] to account for the three-dimensional effect. The existence and uniqueness of a viscosity solution to an equation similar to (1.3) was proved in [19] by the viscosity solution method in [6] for an $\mathbb{R}^2$-valued function ρ, since a scalar stream function ψ can be found in that case such that $\rho = \nabla^\perp \psi$. Such a technique obviously is not applicable here. To our knowledge, the general existence to (1.3) is still open except for the stationary solutions studied in [4].

To draw an analogy with (1.1), we consider a modification of the above equation:

$$\begin{cases} \partial_t \rho + \operatorname{div}(u|\rho|) = 0, \quad (t,x) \in (0,\infty) \times \mathbb{R}^2, \\ u = \nabla \triangle^{-1} \rho, \\ \rho|_{t=0} = \rho_0. \end{cases} \tag{1.4}$$

Notice that when $\rho_0 \geq 0$, (1.4) is the same as (1.1).

This paper consists of two main parts. In the first part, we study the global existence of weak solutions to (1.1) with general Radon measure as initial data under the condition that $\cos\theta \neq 0$ and obtain more general results than those given in [21]. Without loss of generality, we restrict ourselves to the case $\cos\theta > 0$; the results for $\cos\theta < 0$ can be similarly obtained. In the second part of the paper, we present an existence result to (1.4) in one space dimension.

To introduce our main results, let us examine the general procedure on proving the global existence of weak solutions. As in [9], the first step is to construct the approximate solution sequences. For simplicity, let us define the following cut-off function:

$$T_\epsilon(\xi) := \begin{cases} \xi, & \xi \geq -1/\epsilon, \\ -1/\epsilon, & \xi \leq -1/\epsilon. \end{cases} \tag{1.5}$$

We study the approximate solution sequence to (1.1) constructed by the equations

$$\begin{cases} \partial_t \rho_\epsilon + u_\epsilon \nabla \rho_\epsilon = -\cos\theta T_\epsilon(\rho_\epsilon)\rho_\epsilon, \quad (t,x) \in \mathbb{R}^+ \times \mathbb{R}^2, \\ u_\epsilon = M(\theta)\nabla \triangle^{-1}\rho_\epsilon, \\ \rho_\epsilon|_{t=0} = \rho_{0,\epsilon}, \end{cases} \tag{1.6}$$

where $\rho_{0,\epsilon} = (\rho_0 \chi_\epsilon) * j_\epsilon$, $\chi_\epsilon(x) = \chi(\frac{x}{\epsilon})$, $\chi \in C_c^\infty(\mathbb{R}^2)$, $\chi(x) = \begin{cases} 1, & |x| \leq 1, \\ 0, & |x| \geq 2, \end{cases}$ and $j_\epsilon(x)$ is the standard Friedrich's mollifier with $\mathrm{supp} j_\epsilon \subset B_\epsilon(0)$.

Let $S_\epsilon(\xi) = |\xi| * j_\epsilon$. The approximate solution sequence to (1.4) may be defined by the following equation:

$$\begin{cases} \partial_t \rho_\epsilon + \mathrm{div}(u_\epsilon S_\epsilon(\rho_\epsilon)) = \epsilon \triangle \rho_\epsilon, \quad (t,x) \in \mathbb{R}^+ \times \mathbb{R}^2, \\ u_\epsilon = \nabla \triangle^{-1}\rho_\epsilon, \\ \rho_\epsilon|_{t=0} = (\rho_0 \chi_\epsilon) * j_\epsilon. \end{cases} \tag{1.7}$$

Now, a main result of this paper can be stated in the following theorem.

THEOREM 1.1. *Let $\rho_0 \in \mathcal{M}(\mathbb{R}^2)$ and $\cos\theta > 0$. Then there exist a subsequence of $\{\rho_\epsilon, u_\epsilon\}$ constructed by (1.6) (still denoted by $\{\rho_\epsilon, u_\epsilon\}$ for convenience), functions $\rho \in L^q_{loc}(\mathbb{R}^+ \times \mathbb{R}^2) \cap L^\infty(\mathbb{R}^+, L^1(\mathbb{R}^2))$ and $u \in L^q_{loc}(\mathbb{R}^+, W^{1,q}_{loc}(\mathbb{R}^2))$ for any $q < 2$, and a positive Radon measure $\mu \in \mathcal{M}^+(\mathbb{R}^+ \times \mathbb{R}^2)$ such that the following hold:*

1. *The following convergence properties and estimates hold:*

$$\rho_\epsilon \rightharpoonup \rho \quad \text{weakly in} \quad L^q_{loc}(\mathbb{R}^+ \times \mathbb{R}^2), \tag{1.8}$$

$$u_\epsilon \rightharpoonup u \quad \text{weakly in} \quad L^q_{loc}(\mathbb{R}^+, W^{1,q}_{loc}(\mathbb{R}^2)), \tag{1.9}$$

$$\left(\rho_\epsilon + \frac{1}{\epsilon}\right)\rho_\epsilon 1_{\rho_\epsilon \leq -\frac{1}{\epsilon}} \rightharpoonup \mu \quad \text{in the sense of} \quad \mathcal{M}(\mathbb{R}^+ \times \mathbb{R}^2), \tag{1.10}$$

$$\int_0^\infty \int_{\mathbb{R}^2} d\mu \leq \int_{\mathbb{R}^2} |\, d\rho_0(x)|. \tag{1.11}$$

2. *The following decay estimates hold:*

$$\rho(t,x) \leq \frac{\cos\theta}{t} \quad \text{for a.e.} \quad (t,x) \in \mathbb{R}^+ \times \mathbb{R}^2. \tag{1.12}$$

3. *For all test functions $\varphi \in C_c^\infty([0,\infty) \times \mathbb{R}^2)$, there holds*

$$\int_0^\infty \int_{\mathbb{R}^2} (\rho \partial_t \varphi + \rho u \nabla \varphi + \mu \varphi)\, dx\, dt + \int_{\mathbb{R}^2} \varphi(0,x)\rho_0\, dx = 0 \tag{1.13}$$

and

$$u = M(\theta)\nabla\triangle^{-1}\rho. \tag{1.14}$$

DEFINITION 1.1. *We call* (ρ, u, μ) *the measure-valued solution to* (1.1) *if* (ρ, u, μ) *satisfies* (1.13) *and* (1.14).

A similar definition was used by Alexandre and Villani in the study of the Boltzmann equation without Grad's angular assumption to the cross section [1]. There, $f(t, x, v)$ was defined as a renormalized solution to the Boltzmann equation with a defect measure $\mu(t, x, v)$ if, for all nonlinearity $\beta \in C^2(\mathbb{R}^+, \mathbb{R}^+)$ satisfying $\beta(0) = 0$, $0 < \beta'(f) \le \frac{C}{1+f}$, $\beta''(f) < 0$, there holds

$$\partial_t \beta(f) + v \cdot \nabla\beta(f) = \beta'(f)Q(f, f) + \mu$$

in the sense of distributions. One may check [1] for more details.

When $\rho_0 \in \mathcal{M}^+(\mathbb{R}^2)$, we have the following improvement of Theorem 1.1.

COROLLARY 1.1. *Let* $0 < \cos\theta < 1$ *and* $\rho_0 \in \mathcal{M}^+(\mathbb{R}^2)$. *Then* $\mu = 0$ *in* (1.13), *and* $\rho \in L^2((0, \infty) \times \mathbb{R}^2)$. *Moreover,*

$$\|\rho\|_{L^2((0,\infty)\times\mathbb{R}^2)} \le C\rho_0(\mathbb{R}^2). \tag{1.15}$$

If $0 \le \rho_0 \in L^p(\mathbb{R}^2)$ *for* $1 < p < \infty$, *we have the following better estimate for* ρ*:*

$$\int_{\mathbb{R}^2} \rho^p(T, x)\,dx + (p-1)\cos\theta \int_0^T \int_{\mathbb{R}^2} \rho^{p+1}(t, x)\,dx\,dt \le \int_{\mathbb{R}^2} \rho_0^p(x)\,dx \quad \text{for a.e. } T \in \mathbb{R}^+. \tag{1.16}$$

Remark 1.1. In comparison with the measure-valued solutions to (1.2) in [8] and that of the one-dimensional two-component Vlasov–Poisson equations in [27, 28], the measure-valued solution to (1.1) here is much more explicit and closer to the distributional weak solution of (1.1). By (1.10), if ρ_0 is a sign-changing Radon measure, μ may not be 0 even if $\{\rho_\epsilon\}$ strongly converges to ρ in $L^q_{\text{loc}}(\mathbb{R}^+ \times \mathbb{R}^2)$ for any $q < 2$. Moreover, from the corollary, we can see that if $\cos\theta > 0$ and ρ_0 is a positive Radon measure, then the approximate solutions satisfy $\rho_\epsilon(t, x) \ge 0$ for all $(t, x) \in \mathbb{R}^+ \times \mathbb{R}^2$, which in turn implies, by the definition of μ, that $\mu = 0$. Thus, (ρ, u) is the classical distributional weak solution to (1.1).

Remark 1.2. If $\cos\theta > 0$ and $0 \le \rho_0(x) \in L^\infty_{\text{comp}}(\mathbb{R}^2)$, following exactly the same procedure as that in [21] and [33], we can prove the uniqueness of the weak solution in the above corollary.

Remark 1.3. We can replace the second equation in (1.1) by $u = -M(\theta)\nabla(-\lambda^2\triangle + 1)^{-1}\rho$ in correspondence to the equations derived in [4]. Step-by-step modifications of the proofs given here will yield similar results for such equations as those in Theorem 1.1.

In one space dimension, (1.4) takes on the following form:

$$\begin{cases} \partial_t \rho + \partial_x(u|\rho|) = 0, \quad (t, x) \in (0, \infty) \times \mathbb{R}, \\ u = \int_{-\infty}^x \rho(t, y)\,dy, \\ \rho|_{t=0} = \rho_0. \end{cases} \tag{1.17}$$

Then we have the following existence result to (1.17).

THEOREM 1.2. *Let* $\rho_0 \in BV(\mathbb{R})$. *Then the weak limit* (ρ, u) *obtained by the vanishing viscosity of* (1.7) *satisfies* (1.17) *in the sense of distributions, and*

$\rho \in C([0,T], L^p(K)) \cap L^\infty([0,T], BV(\mathbb{R}))$, $u \in L^\infty([0,T], W^{1,p}(\mathbb{R}))$ *for any* $T < \infty$, $2 < p < \infty$, *and any compact subset* K *of* $\mathbb{R}$. *Furthermore,*

$$\int_{\mathbb{R}} |\rho(t,x)|\,dx \le \int_{\mathbb{R}} |\rho_0|\,dx, \quad \int_{\mathbb{R}} |\,d\rho(t,x)| \le \int_{\mathbb{R}} |\,d\rho_0|, \tag{1.18}$$

where $\int_{\mathbb{R}} |\,d\rho|$ *denotes the total variation of* ρ, *if* $\operatorname{supp}\rho_0 \subset B_r(0)$, $\operatorname{supp}\rho(t,\cdot) \subset \{x : |x| \le r + Mt\}$ *with* M *the* L^∞ *bound of* u.

Remark 1.4. (1) An explicit solution to (1.17) may be constructed: let ρ_0 be defined by $\rho_0(x) = 1$ in $(0,1)$, $\rho_0(x) = -1$ in $(1,2)$, and $\rho_0(x) = 0$ elsewhere; then a global weak solution to (1.17) may be defined by

$$\rho(t,x) = \begin{cases} \frac{1}{1+t}, & x \in (0,1), \\ \frac{-1}{1+t}, & x \in (1,2), \\ 0 & \text{otherwise.} \end{cases}$$

Here, $x = 1$ is the shock line of this solution. Naturally, this leads to a conjecture that shocks will be formed for $d = 2$ when ρ_0 changes sign. In general we cannot prove $|\rho_\epsilon| \rightharpoonup m = |\rho|$. Nevertheless, we can prove that

$$\partial_t \rho + \operatorname{div}(um) = 0, \quad u = \nabla \triangle^{-1} \rho$$

holds in the sense of distributions and $m = |\rho|$ for almost all (t,x) in a subset of $\mathbb{R}^+ \times \mathbb{R}^2$; see Proposition 3.1 for more details.

(2) With $d = 2$, to get the uniform $L^\infty([0,T], L^1(\mathbb{R}^2))$ estimate for $\{\partial_x \rho_\epsilon\}$, the solution sequence to (1.7), we need the uniform L^∞ estimate for ∇u_ϵ (see the proof of Lemma 3.2 for details). But with ρ_ϵ being uniformly bounded, we cannot get the desired estimate for $\nabla \otimes u_\epsilon = \nabla \otimes \nabla \triangle^{-1} \rho_\epsilon$ by the singular integral operator theory [31]. However, in the case of $d = 1$, by the second equation of (1.17), we find $\partial_x u_\epsilon = \rho_\epsilon$, which gives the desired estimate for $\partial_x u_\epsilon$.

Remark 1.5. Again, one may replace the second equation of (1.17) by $u = -\partial_x(-\lambda^2 \partial_{xx} + 1)^{-1} \rho$ and, using the same type of arguments as those in Theorem 1.2, prove similar results.

We now introduce some notation that will be used throughout the paper. We let $B_r(0) = \{x : |x| \le r\}$ and denote by $\mathcal{M}(\Omega)$ the bounded Radon measure space on Ω, by $\mathcal{M}^+(\Omega)$ the bounded positive Radon measure space on Ω, and $BV(\mathbb{R}) = \{f : f \in L^1(\mathbb{R}), \partial_x f \in \mathcal{M}(\mathbb{R})\}$. We use $C(a,b,\ldots)$ as a uniform constant which only depends on the listed variables and may change from line to line.

The proofs of the above theorems are given in later sections.

2. Proof of Theorem 1.1. Now let us first prove the global existence of solutions to (1.6). For convenience, we omit the subscript ϵ in the approximate solution sequence $\{(\rho_\epsilon, u_\epsilon)\}$ in the following lemma.

LEMMA 2.1 (solution of (1.6) with smooth data). *For* $\rho_0 \in C_c^\infty(\mathbb{R}^2)$, (1.6) *has a global strong solution* (ρ, u) *such that* $\rho \in L^\infty([0,T], W^{1,p}(\mathbb{R}^2))$, $\nabla u \in L^\infty([0,T], W^{1,p}(\mathbb{R}^2))$ *for any* $p > 1$, $T < \infty$, *and*

$$\|\rho(t,\cdot)\|_{L^1} \le \|\rho_0\|_{L^1} \quad \textit{and} \quad \rho(t,x) \le \frac{\cos\theta}{t}, \qquad t > 0. \tag{2.1}$$

Proof. 1. (Blow-up principle.) Following the standard argument for a nonlinear hyperbolic equation, we can prove the local existence of solution (ρ, u) to (1.6) with

smooth data such that $\rho, \nabla u \in L^\infty([0,T], W^{1,p}(\mathbb{R}^2))$ for some positive constant T and any $p < \infty$. Now, let T^* be the lifespan of the solution (ρ, u) to (1.6). We are going to show that if $T^* < \infty$,

$$\lim_{t \to T^*} \|\rho(t,\cdot)\|_{L^\infty} = \infty. \tag{2.2}$$

In fact, for any even positive number p, it follows from (1.6) that

$$\begin{aligned} \partial_t(\partial_{x_i}\rho)^p + u\nabla(\partial_{x_i}\rho)^p + p\partial_{x_i}u\nabla\rho(\partial_{x_i}\rho)^{p-1} \\ = -p\cos\theta\partial_{x_i}(T_\epsilon(\rho)\rho)(\partial_{x_i}\rho)^{p-1}. \end{aligned} \tag{2.3}$$

Noticing that $\mathrm{div} u = \cos\theta\rho$, $|\partial_{x_i}(T_\epsilon(\rho)\rho)| \le 2|\rho||\partial_{x_i}\rho|$, integrating the above equation over $\mathbb{R}^2$, and using integration by parts, we get

$$\frac{d}{dt}\int_{\mathbb{R}^2} |\partial_{x_i}\rho|^p\,dx \le ((2p+1)\cos\theta\|\rho\|_{L^\infty} + p\|\partial_{x_i}u\|_{L^\infty})\int_{\mathbb{R}^2}|\nabla\rho|^p\,dx. \tag{2.4}$$

Let us take $\chi(\xi) \in C_c^\infty(\mathbb{R}^2), \chi(D)$ the corresponding pseudodifferential operator with symbol $\chi(\xi)$; then by singular integrals theory [31, 32], we have

$$\|\chi(D)\nabla\otimes\nabla\triangle^{-1}\rho\|_{L^\infty} \le C\|\nabla\otimes\nabla\triangle^{-1}\rho\|_{L^p} \le C\|\rho\|_{L^p}. \tag{2.5}$$

While by Lemma B.1.C of [32], for all $p > 2$, we find

$$\|(1-\chi(D))\nabla\otimes\nabla\triangle^{-1}\rho\|_{L^\infty} \le C\|\rho\|_{L^\infty}\left(1 + \log\frac{\|\rho\|_{W^{1,p}}}{\|\rho\|_{L^\infty}}\right). \tag{2.6}$$

Summing up the second equation of (1.6) and inequalities (2.5) and (2.6), we find

$$\|\nabla u\|_{L^\infty} \le C\left\{\|\rho\|_{L^\infty}\left(1 + \log\frac{\|\rho\|_{W^{1,p}}}{\|\rho\|_{L^\infty}}\right) + \|\rho\|_{L^p}\right\}. \tag{2.7}$$

A simple interpolation result gives us

$$\|\rho\|_{L^p} \le \|\rho\|_{L^\infty}^{\frac{p-1}{p}}\|\rho\|_{L^1}^{\frac{1}{p}}. \tag{2.8}$$

Summing up (2.4) and (2.7)–(2.8) we obtain

$$\frac{1}{2}\frac{d}{dt}\|\nabla\rho(t,\cdot)\|_{L^p}^p \le C\left\{\|\rho\|_{L^\infty}\left(1 + \log\frac{\|\nabla\rho\|_{L^p}}{\|\rho\|_{L^\infty}}\right) + \|\rho\|_{L^1}\right\}\|\nabla\rho\|_{L^p}^p. \tag{2.9}$$

Then the Gronwall inequality yields that

$$\|\nabla\rho(t,\cdot)\|_{L^p} \le C(T, \|\rho\|_{L^1}, \|\rho\|_{L^\infty})\|\nabla\rho_0\|_{L^p}. \tag{2.10}$$

On the other hand, by multiplying $\mathrm{sign}\rho$ on both sides of (1.6), we find by (1.5) that

$$\partial_t|\rho| + \mathrm{div}(u|\rho|) = \cos\theta(\rho - T_\epsilon(\rho))|\rho| \le 0. \tag{2.11}$$

Integrating (2.11) over $\mathbb{R}^2$, we get the first inequality of (2.1). Summing up (2.1) and (2.10), we complete the proof of the claim (2.2).

2. (Estimate of $\|\rho\|_{L^\infty}$.) By (2.7) and the classical theory on ordinary differential equations, the equation

$$\begin{cases} \frac{d\Phi_t(x)}{dt} = u(t, \Phi_t(x)), \\ \Phi_t(x)|_{t=0} = x \end{cases} \tag{2.12}$$

has a unique solution $\Phi_t(x) \in C([0,T^*)\times\mathbb{R}^2)$, and $\partial_x\Phi_t(x) \in L^\infty([0,T]\times\mathbb{R}^2)$ for any $T<T^*$. Then, by the first equation of (1.6), we have

$$\frac{d\rho(t,\Phi_t(x))}{dt} \le 0,$$

which implies that

$$\rho(t,\cdot) \le \|\rho_0\|_{L^\infty}. \tag{2.13}$$

This together with (1.5) shows that

$$\|T_\epsilon(\rho)\|_{L^\infty} \le \max\left\{\frac{1}{\epsilon}, \|\rho_0\|_{L^\infty}\right\}. \tag{2.14}$$

Thus by the first equation of (1.6) and by (2.12), we have

$$\frac{d|\rho|(t,\Phi_t(x))}{dt} \le \cos\theta \max\left\{\frac{1}{\epsilon}, \|\rho_0\|_{L^\infty}\right\} |\rho|(t,\Phi_t(x)),$$

which together with the Gronwall inequality yields that

$$\|\rho\|_{L^\infty} \le \exp\left(\cos\theta \max\left\{\frac{1}{\epsilon}, \|\rho_0\|_{L^\infty}\right\} t\right) \|\rho_0\|_{L^\infty}. \tag{2.15}$$

Summing up (2.2) and (2.15), we get the global existence of strong solutions to (1.6) with smooth initial data.

3. (Decay estimate.) By the first equation of (1.6) and by (2.12), we have

$$\frac{d\rho(t,\Phi_t(x))}{dt} = -\cos\theta (T_\epsilon(\rho)\rho)(t,\Phi_t(x)), \tag{2.16}$$

which implies that if $\rho_0(x) \le 0$, then $\rho(t,\Phi_t(x)) \le 0$, and if $\rho_0(x) \ge 0$, then $\rho(t,\Phi_t(x)) \ge 0$. So to prove the one-sided decay estimate (2.1), we need to consider only the points where $\rho_0(x) > 0$. By (1.5), $T_\epsilon(\rho(t,\Phi_t(x)) = \rho(t,\Phi_t(x))$. Solving (2.16), we get

$$\rho(t,\Phi_t(x)) = \frac{\cos\theta\rho_0(x)}{1+t\rho_0(x)} < \frac{\cos\theta}{t}, \quad t>0. \tag{2.17}$$

Summing up the above, we get the second inequality of (2.1). This completes the proof of the lemma. □

Next let us get the key uniform space-time estimate for the approximate solution sequence $\{\rho_\epsilon\}$ constructed in Lemma 2.1.

LEMMA 2.2 ($L^{1+\alpha}$ estimate). *Let $\rho_{0,\epsilon} \in L^1(\mathbb{R}^2), \alpha \in (0,1), T, R > 0$. Then for the solutions $(\rho_\epsilon, u_\epsilon)_{\epsilon>0}$ of (1.6), there holds the estimate*

$$\int_0^T \int_{|x|\le R} |\rho_\epsilon|^{1+\alpha}\, dx\, dt \le C_{\alpha,T,R}, \tag{2.18}$$

where the constant $C_{\alpha,T,R}$ depends only on the L^1 norm of $\rho_{0,\epsilon}$ and the listed variables.

Proof. 1. (Elementary estimate.) We first assume $\alpha = d_2/d_1 \in (0,1/2)$, where d_1 and d_2 are odd positive integers. Let $\chi \in C_c^\infty(\mathbb{R}^2), \chi \ge 0$ and $\chi = 1$ on $\{x||x| \le R\}$,

with $\operatorname{supp}\chi \subset \{x||x| \le R+1\}$. Set $\eta(\xi) = \alpha\int_0^\xi \max(1,|s|)^{\alpha-1}\,ds$ for $\xi \in \mathbb{R}^1$ such that $\eta'(\xi) = \alpha\max(1,|\xi|)^{\alpha-1}$. Multiplying (1.6) by $\chi\eta'(\rho_\epsilon)$, integrating the resulting identity over $[0,T]\times\mathbb{R}^2$, and performing integration by parts several times, we obtain

$$\begin{aligned}(2.19)\quad \cos\theta\int_0^T\int_{\mathbb{R}^2}\chi(\rho_\epsilon\eta(\rho_\epsilon)-\rho_\epsilon T_\epsilon(\rho_\epsilon)\eta'(\rho_\epsilon))\,dx\,dt\\ =\int_{\mathbb{R}^2}\chi\eta(\rho_\epsilon)\,dx|_0^T-\int_0^T\int_{\mathbb{R}^2}\nabla\chi u_\epsilon\eta(\rho_\epsilon)\,dx\,ds.\end{aligned}$$

First, by the second equation of (1.6), for all test functions $\phi(x),\psi(x)\in C_c^\infty(\mathbb{R}^2)$, we have

$$\begin{aligned}\phi(x)\psi(x)u_\epsilon(t,x) &= M(\theta)\phi(x)\int_{\mathbb{R}^2}(\psi(x)-\psi(y))\frac{x-y}{|x-y|^2}\rho_\epsilon(t,y)\,dy\\ &\quad + M(\theta)\int_{\mathbb{R}^2}\phi(x)\psi(y)\frac{x-y}{|x-y|^2}\rho_\epsilon(t,y)\,dy.\end{aligned}\tag{2.20}$$

Notice that $\phi(x)\psi(y)\frac{x-y}{|x-y|^2} = \phi(x)\psi(y)\zeta(|x-y|)\frac{x-y}{|x-y|^2}$, where $\zeta(z)\in C_c^\infty(\mathbb{R})$ with $\zeta(z)=1$ for $z\in\operatorname{supp}\phi+\operatorname{supp}\psi$. And trivially $\zeta(z)\frac{z}{|z|^2}\in L^p(\mathbb{R}^2)$ for all $p<2$, by the Hausdorff–Young inequality to the second term of (2.20), we get

$$\|\phi\psi u_\epsilon(t,\cdot)\|_{L^p} \le (\sup|\nabla\psi|\|\phi\|_{L^p}+c_{\phi,\psi})\|\rho_\epsilon(t,\cdot)\|_{L^1}\quad\text{for all } 1\le p<2.$$

In particular, by taking $\phi(x)=\psi(x)=1$ for $|x|\le R+1$, we get

$$\left(\int_{|x|\le R+1}|u_\epsilon(t,x)|^p\,dx\right)^{\frac{1}{p}} \le C_R\|\rho_{0,\epsilon}\|_{L^1}\quad\text{for all } 1<p<2.\tag{2.21}$$

Note that $\alpha<\frac12$, and thus by (2.1) and (2.21), we have

$$\begin{aligned}\left|\int_0^T\int_{\mathbb{R}^2}\nabla\chi u_\epsilon\eta(\rho_\epsilon)\,dx\,ds\right| &\le \int_0^T\left(\int_{|x|\le R+1}|u_\epsilon|^{\frac{1}{1-\alpha}}\,dx\right)^{1-\alpha}\left(\int_{|x|\le R+1}|\rho_\epsilon|\,dx\right)^{\alpha}dt\\ &\le C_1(R,\|\rho_{0,\epsilon}\|_{L^1})\,.\end{aligned}\tag{2.22}$$

By (2.1) and the definition of η, we have

$$\begin{aligned}&\left|\int_{\mathbb{R}^2}\chi\eta(\rho_\epsilon)(T,x)\,dx\right|\\ &\le\int_{|\rho_\epsilon|\ge1}\chi(\alpha+|\rho_\epsilon|^\alpha)(T,x)\,dx+\alpha\int_{|\rho_\epsilon|\le1}\chi(x)|\rho_\epsilon|\,dx\\ &\le 2\pi\alpha(R+1)^2+\left(\int_{|x|\le R+1}|\rho_\epsilon|(T,x)\,dx\right)^{\alpha}(2\pi(R+1)^2)^{1-\alpha}+\alpha2\pi(R+1)^2\\ &\le C_2(\alpha,R,\|\rho_{0,\epsilon}\|_{L^1}).\end{aligned}\tag{2.23}$$

Finally it follows from the definition of α and η that

$$\begin{aligned}&\int_0^T\int_{\mathbb{R}^2}\chi(x)(\rho_\epsilon\eta(\rho_\epsilon)-\rho_\epsilon T_\epsilon(\rho_\epsilon)\eta'(\rho_\epsilon))\,dx\,dt\\ &=\int_0^T\int_{\mathbb{R}^2}1_{|\rho_\epsilon|\ge1}\chi\left((1-\alpha)\rho_\epsilon^{1+\alpha}+\alpha(\rho_\epsilon^{1+\alpha}-T_\epsilon(\rho_\epsilon)\rho_\epsilon^\alpha)+\alpha\rho_\epsilon\right)dx\\ &\ge\int_0^T\int_{\mathbb{R}^2}1_{|\rho_\epsilon|\ge1}\chi((1-\alpha)\rho_\epsilon^{1+\alpha}+\alpha\rho_\epsilon)\,dx,\end{aligned}\tag{2.24}$$

where $1_{|\rho_\epsilon|\geq 1}$ is the characteristic function on the set $\{(t,x) : |\rho_\epsilon(t,x)| \geq 1\}$.

Summing up (2.19)–(2.24), we find

$$\int_0^T \int_{|\rho_\epsilon|\geq 1} \chi \rho_\epsilon^{1+\alpha}\, dx\, dt \leq \frac{1}{1-\alpha}\left(\alpha \int_{\mathbb{R}^2} \chi|\rho_\epsilon|\, dx + C_1 + C_2\right) \tag{2.25}$$

for all $\alpha = d_2/d_1 \in (0, 1/2)$.

2. (Inductive step 1.) Next, we are going to show by the bootstrap method that (2.25) holds for all $\alpha \in (0,1)$. First, let us take $\alpha = d_2/d_1 \in (0, 5/6)$ with d_1, d_2 positive odd integers. In particular, due to the arbitrariness of R, by interpolation, and by (2.25), we have

$$\int_0^T \int_{|x|\leq R+1} |\rho_\epsilon|^{p_1}\, dx\, dt \leq C(R, T, \|\rho_{0,\epsilon}) \quad \text{for all } p_1 < \frac{3}{2}. \tag{2.26}$$

On the other hand, again by (2.20) and the Hausdorff–Young inequality, we have

$$\begin{aligned} &\int_0^T \left(\int_{\mathbb{R}^2} |\phi\psi u_\epsilon(t,\cdot)|^{q_1}\, dx\right)^{p_1/q_1} dt \\ &\leq (\sup|\nabla\psi| \|\rho_\epsilon(t,\cdot)\|_{L^1} \|\phi\|_{L^{q_1}})^{p_1}\, T + C_{\phi,\psi} \int_0^T \|\psi\rho_\epsilon(t,\cdot)\|_{L^{p_1}}^{p_1}\, dt, \end{aligned} \tag{2.27}$$

with $\frac{1}{q_1} = \frac{1}{p_1} + \frac{1}{p} - 1$ for any $p < 2$. This implies that

$$\int_0^T \left(\int_{|x|\leq R+1} |u_\epsilon(t,x)|^{q_1}\, dx\right)^{\frac{p_1}{q_1}} dt \leq C(R, T, \|\rho_{0,\epsilon}\|_{L^1}), \tag{2.28}$$

with the same p_1 and q_1. Thus by the Hölder inequality, we have

$$\begin{aligned} \left|\int_0^T \int_{\mathbb{R}^2} \nabla\chi u_\epsilon \eta(\rho_\epsilon)\, dx\, ds\right| &\leq \int_0^T \left(\int_{|x|\leq R+1} |u_\epsilon|^{\frac{1}{1-\alpha}}\, dx\right)^{1-\alpha} \left(\int_{|x|\leq R+1} |\rho_\epsilon|\, dx\right)^{\alpha} dt \\ &\leq \|\rho_{0,\epsilon}\|_{L^1}^{\alpha} \int_0^T \left(\int_{|x|\leq R+1} |u_\epsilon|^{\frac{1}{1-\alpha}}\, dx\right)^{1-\alpha} dt \\ &\leq C_3(\chi, T, \|\rho_{0,\epsilon}\|_{L^1}), \end{aligned} \tag{2.29}$$

with $1 - \alpha \geq \frac{1}{p_1} + \frac{1}{p} - 1$. As $p_1 < \frac{3}{2}, p < 2$, we can always take $\alpha = \frac{d_2}{d_1} \in (0, \frac{5}{6})$ such that $1 - \alpha \geq \frac{1}{p_1} + \frac{1}{p} - 1$.

Summing up (2.19), (2.23), (2.24), and (2.29), we then find

$$\int_0^T \int_{|\rho_\epsilon|\geq 1} \chi \rho_\epsilon^{1+\alpha}\, dx\, dt \leq \frac{1}{1-\alpha}\left(\int_{\mathbb{R}^2} \chi|\rho_\epsilon|\, dx + C_2 + C_3\right). \tag{2.30}$$

This together with interpolation theory implies

$$\int_0^T \int_{|x\leq R+1} |\rho_\epsilon|^{p_2}\, dx\, dt \leq C(\alpha, R, T, \|\rho_{0,\epsilon}\|_{L^1}) \quad \text{for all } p_2 < \frac{11}{6}. \tag{2.31}$$

3. (Induction.) Inductively, let us assume $\alpha \in (0, \alpha_n)$, with $\frac{1}{2} < \alpha_n < 1$, and set $\bar{p}_n = 1 + \alpha_n$; there holds the estimate (2.18). Then by (2.27) and the similar proof of (2.28), we have

$$\int_0^T \left(\int_{|x| \le R+1} |\chi u_\epsilon|^{q_{n+1}} \, dx \right)^{\frac{p_n}{q_{n+1}}} dt \le C(\chi, T, \|\rho_{0,\epsilon}\|_{L^1}), \tag{2.32}$$

with $\frac{1}{q_{n+1}} = \frac{1}{p_n} - \frac{1}{2}$ and $p_n < \bar{p}_n$. Thus by the similar proof of (2.28), we have

$$\begin{aligned} \left| \int_0^T \int_{\mathbb{R}^2} \nabla \chi u_\epsilon \eta(\rho_\epsilon) \, dx \, dt \right| &\le \|\rho_{0,\epsilon}\|_{L^1}^{\alpha} \int_0^T \left(\int_{|x| \le R+1} |u_\epsilon|^{\frac{1}{1-\alpha}} \, dx \right)^{1-\alpha} dt \\ &\le C(\chi, T, \|\rho_{0,\epsilon}\|_{L^1}), \end{aligned} \tag{2.33}$$

with $1 - \alpha \ge \frac{1}{p_n} - \frac{1}{2}$. As $p_n < 1 + \alpha_n$, we can always take $\alpha = \frac{d_2}{d_1} \in (0, \alpha_{n+1})$ with d_1, d_2 being positive odd integers, and $\alpha_{n+1} = \frac{1+3\alpha_n}{2(1+\alpha_n)}$ such that $1 - \alpha \ge \frac{1}{p_n} - \frac{1}{2}$. Then, summing up (2.19), (2.23), (2.24), and (2.33), we find there holds (2.29) with $\alpha \in (0, \alpha_{n+1})$, which implies that

$$\int_0^T \int_{|x| \le R+1} |\rho_\epsilon|^{p_{n+1}} \, dx \, dt \le C(\alpha, R, T, \|\rho_{0,\epsilon}\|_{L^1}) \quad \text{for all } p_{n+1} < 1 + \alpha_{n+1}. \tag{2.34}$$

Notice by the definition of α_n we have $\alpha_{n+1} > \alpha_n$ if $1/2 < \alpha_n < 1$. Thus, the limit $\lim_{n\to\infty} \alpha_n$ exists. Moreover, by the inductive formula $\alpha_{n+1} = \frac{1+3\alpha_n}{2(1+\alpha_n)}$, we have

$$\lim_{n \to \infty} \alpha_n = 1. \tag{2.35}$$

Summing up (2.34) and (2.35), we complete the proof of Lemma 2.2. □

When $\rho_{0,\epsilon} \ge 0$ and $0 < \cos\theta < 1$, we can have the following improved estimates for the approximate solution sequence $\{\rho_\epsilon\}$.

LEMMA 2.3 (L^{p+1} estimate). *Let $\rho_{0,\epsilon} \ge 0$ and $0 < \cos\theta < 1$. The solution sequence $\{\rho_\epsilon\}$ then satisfies*

$$\int_0^T \int_{\mathbb{R}^2} \rho_\epsilon^2 \, dx \, dt \le C \|\rho_{0,\epsilon}\|_{L^1}, \tag{2.36}$$

$$\int_{\mathbb{R}^2} \rho_\epsilon^p(T, x) \, dx + (p-1)\cos\theta \int_0^T \int_{\mathbb{R}^2} \rho_\epsilon^{p+1} \, dx \, dt = \int_{\mathbb{R}^2} \rho_{0,\epsilon}^p \, dx \tag{2.37}$$

for $1 < p < \infty$.

Proof. First, by the first equation of (1.6), we know that if $\rho_{0,\epsilon} \ge 0$, then $\rho_\epsilon(t, x) \ge 0$ for $(t, x) \in \mathbb{R}^+ \times \mathbb{R}^2$. Then, by the definition of $T_\epsilon(\xi)$ in (1.5), we have $T_\epsilon(\rho_\epsilon) = \rho_\epsilon$ and

$$\partial_t \rho_\epsilon + u_\epsilon \nabla \rho_\epsilon = -\cos\theta \rho_\epsilon^2. \tag{2.38}$$

Thus by (2.12) and (2.17), we have

$$\begin{aligned} \det\left(\frac{\partial \Phi_t^\epsilon(x)}{\partial x} \right) &= \exp\left(\int_0^t \mathrm{div} u_\epsilon(s, \Phi_s^\epsilon(x)) \, ds \right) \\ &= \exp\left(\cos\theta \int_0^t \rho_\epsilon(s, \Phi_s^\epsilon(x)) \, ds \right) \\ &= \exp\left(\cos\theta \int_0^t \frac{\rho_{0,\epsilon}(x)}{1 + s\rho_{0,\epsilon}(x)} \, ds \right) \\ &= (1 + t\rho_{0,\epsilon})^{\cos\theta}. \end{aligned} \tag{2.39}$$

Moreover, by (2.12) and (2.17), we can write the solution to (1.6) in the following form:

$$\rho_\epsilon(t,y)=\frac{\cos\theta\rho_{0,\epsilon}((\Phi_s^\epsilon)^{-1}(y))}{1+t\rho_{0,\epsilon}((\Phi_s^\epsilon)^{-1}(y))}. \tag{2.40}$$

Thus, summing up (2.39) and (2.40), we have for any $T<\infty$ and $0<\cos\theta<1$ that

$$\begin{aligned}\int_0^T\int_{\mathbb{R}^2}\rho_\epsilon^2\,dy\,dt&=\int_0^T\int_{\mathbb{R}^2}\left(\frac{\cos\theta\rho_{0,\epsilon}((\Phi_s^\epsilon)^{-1}(y))}{1+t\rho_{0,\epsilon}((\Phi_s^\epsilon)^{-1}(y))}\right)^2dy\,dt\\&=\int_0^T\int_{\mathbb{R}^2}\frac{\cos^2\theta\rho_{0,\epsilon}^2}{(1+t\rho_{0,\epsilon}(x))^{2-\cos\theta}}\,dx\,dt\\&=\frac{\cos^2\theta}{\cos\theta-1}\int_{\mathbb{R}^2}\rho_{0,\epsilon}(1+t\rho_{0,\epsilon})^{\cos\theta-1}|_0^T\,dx\\&\le\frac{1}{1-\cos\theta}\int_{\mathbb{R}^2}\rho_{0,\epsilon}\,dx\le C\|\rho_{0,\epsilon}\|_{L^1},\end{aligned} \tag{2.41}$$

which proves (2.36).

On the other hand, multiplying $p\rho_\epsilon^{p-1}$ with (2.38), we find

$$\partial_t\rho_\epsilon^p+\operatorname{div}(u_\epsilon\rho_\epsilon^p)=\cos\theta(1-p)\rho_\epsilon^{p+1}. \tag{2.42}$$

Integrating (2.42) over $[0,T]\times\mathbb{R}^2$, we get (2.37). This proves the lemma. □

Now, we are in a position to complete the proof of Theorem 1.1.

Proof of Theorem 1.1. First by (2.18), there is a subsequence of $\{\rho_\epsilon\}$, which is still denoted by $\{\rho_\epsilon\}$ for convenience, and some $\rho\in L^q_{\rm loc}(\mathbb{R}^+\times\mathbb{R}^2)$ such that

$$\{\rho_\epsilon\}\rightharpoonup\rho\quad\text{weakly in}\quad L^q([0,T],L^q_{\rm loc}(\mathbb{R}^2))\quad\text{for all } q<2 \tag{2.43}$$

as $\epsilon\to0$. Thus by (2.43) and the first equation of (2.1), for all $\varphi\in C_c^\infty(\mathbb{R}^+\times\mathbb{R}^2)$, there holds

$$\begin{aligned}\left|\int_{\mathbb{R}^+}\int_{\mathbb{R}^2}\varphi\rho\,dx\,dt\right|&=\left|\lim_{\epsilon\to0}\int_{\mathbb{R}^+}\int_{\mathbb{R}^2}\varphi\rho_\epsilon\,dx\,dt\right|\\&\le\sup_{t\in\mathbb{R}^+}\|\rho_\epsilon(t,\cdot)\|_{L^1}\int_0^\infty\|\varphi(t,\cdot)\|_{L^\infty}\,dt\\&\le C\int_0^\infty\|\varphi(t,\cdot)\|_{L^\infty}\,dt,\end{aligned} \tag{2.44}$$

which implies that $\rho\in L^\infty(\mathbb{R}^+,L^1(\mathbb{R}^2))$.

Notice that the first equation of (1.6) can be rewritten as

$$\begin{aligned}\partial_t\rho_\epsilon+\operatorname{div}(u_\epsilon\rho_\epsilon)&=\cos\theta(\rho_\epsilon^2-T_\epsilon(\rho_\epsilon)\rho_\epsilon)\\&=\cos\theta\left(\rho_\epsilon+\frac{1}{\epsilon}\right)1_{\rho_\epsilon\le-\frac{1}{\epsilon}}\rho_\epsilon.\end{aligned} \tag{2.45}$$

By integrating the above equation over $[0,T]\times\mathbb{R}^2$ and using (2.1), we find

$$\begin{aligned}\cos\theta\int_0^T\int_{\mathbb{R}^2}\left(\rho_\epsilon+\frac{1}{\epsilon}\right)\rho_\epsilon1_{\rho_\epsilon\le-\frac{1}{\epsilon}}\,dx\,dt&=\int_{\mathbb{R}^2}\rho_\epsilon(T,x)\,dx-\int_{\mathbb{R}^2}\rho_{0,\epsilon}\,dx\\&\le2\int_{\mathbb{R}^2}|\rho_{0,\epsilon}|\,dx.\end{aligned} \tag{2.46}$$

Thus by [20], up to a subsequence, which we still denote by $\{(\rho_\epsilon + \frac{1}{\epsilon})\rho_\epsilon 1_{\rho_\epsilon \le -\frac{1}{\epsilon}}\}$ for convenience, there exists a positive Radon measure μ on $R \times \mathbb{R}^2$ such that

$$\left(\rho_\epsilon + \frac{1}{\epsilon}\right)\rho_\epsilon 1_{\rho_\epsilon \le -\frac{1}{\epsilon}} \rightharpoonup \mu \quad \text{in the sense of measure as} \quad \epsilon \to 0. \tag{2.47}$$

On the other hand, by (2.1), (2.18), and interpolation theory,

$$\{\rho_\epsilon\} \quad \text{is uniformly bounded in} \quad L^{p_1}([0,T], L^{p_2}_{\text{loc}}(\mathbb{R}^2)), \tag{2.48}$$

with $\frac{1}{p_1} = \frac{\beta}{\infty} + \frac{1-\beta}{q}$, $\frac{1}{p_2} = \frac{\beta}{1} + \frac{1-\beta}{q}$ for any $0 < \beta < 1, 1 < q < 2$. Since $u_\epsilon = M(\theta)\nabla\triangle^{-1}\rho_\epsilon$, by Bessel potential theory [31], we have

$$\{u_\epsilon\} \quad \text{is uniformly bounded in} \quad L^{p_1}([0,T], L^{p_3}_{\text{loc}}(\mathbb{R}^2)), \tag{2.49}$$

with $\frac{1}{p_3} = \frac{1}{p_2} - \frac{1}{2}$. If we take $\beta = 2 - q$ for $\frac{3}{2} < q < 2$, then $\frac{1}{p_1} + \frac{1}{q} = 1$ and $\frac{1}{q} + \frac{1}{p_3} = \frac{1}{2} + \beta < 1$. With these particular choices, and with (2.45), (2.46), we find

$$\{\partial_t \rho_\epsilon\} \quad \text{is uniformly bounded in} \quad L^1\left([0,T], W^{-1,\frac{2}{5-2q}}_{\text{loc}}(\mathbb{R}^2) + L^1(\mathbb{R}^2)\right). \tag{2.50}$$

Moreover by the definition of u_ϵ, (2.48), and Riesz transform theory [31], we have

$$\{\nabla_x u_\epsilon\} \quad \text{is uniformly bounded in} \quad L^{p_1}([0,T], L^{p_2}_{\text{loc}}(\mathbb{R}^2)), \tag{2.51}$$

so that

$$\|u_\epsilon(t,\cdot) - u_\epsilon(t, x+\xi)\|_{L^{p_1}([0,T], L^p(B_R))} \to 0 \quad \text{as} \quad |\xi| \to 0 \tag{2.52}$$

for all $p < p_3$ but close to p_3, so that $\frac{1}{q} + \frac{1}{p} \le 1$.

Thus if we denote u to be the weak limits of $\{u_\epsilon\}$ in $L^{p_1}([0,T], L^{p_3}_{\text{loc}}(\mathbb{R}^2))$, summing up (2.50), (2.52) and using Lemma 5.1 of [23], we find

$$\rho_\epsilon u_\epsilon \rightharpoonup \rho u \quad \text{in the sense of distributions as} \quad \epsilon \to 0. \tag{2.53}$$

While by (2.45) and integration by parts, for any test function $\varphi \in C^\infty_c([0,T] \times \mathbb{R}^2)$, we obviously have

$$\begin{aligned} \int_0^T \int_{\mathbb{R}^2} (\partial_t \varphi \rho_\epsilon + \rho_\epsilon u_\epsilon \nabla \varphi)\, dx\, dt + \cos\theta \int_0^T \int_{\mathbb{R}^2} \varphi \left(\rho_\epsilon + \frac{1}{\epsilon}\right)\rho_\epsilon 1_{\rho_\epsilon \le -\frac{1}{\epsilon}}\, dx\, dt \\ + \int_{\mathbb{R}^2} \varphi \rho_{0,\epsilon}\, dx = 0. \end{aligned} \tag{2.54}$$

Summing up (2.43), (2.47), and (2.53), and taking ϵ to 0 in (2.54), we get (1.13). Moreover, by the second equation of (1.6), it is trivial to get (1.14). By summing up (2.1) and (2.43), we get (1.12). This completes the proof of Theorem 1.1. □

Now let us turn to the proof of Corollary 1.1.

Proof of Corollary 1.1. First by the proof of Lemma 2.1, if ρ_0 is a positive Radon measure, then $\rho_\epsilon(t,x) \ge 0$ for all $(t,x) \in \mathbb{R}^+ \times \mathbb{R}^2$, and thus by (2.47), $\mu(t,x) = 0$. Moreover, by (2.36) and [20], we have

$$\|\rho(t,x)\|_{L^2([0,T]\times\mathbb{R}^2)} \le \underline{\lim}_{\epsilon\to 0}\|\rho_\epsilon(t,x)\|_{L^2([0,T]\times\mathbb{R}^2)} \le C\rho_0(\mathbb{R}^2) \tag{2.55}$$

for all $0 < \cos\theta < 1$. Then (2.55) and Fatou's lemma imply (1.15). For $p > 1$, again by [20] and (2.37), we have

$$\begin{aligned}\int_{\mathbb{R}^2} \rho^p(T,x)\,dx + (p-1)\cos\theta \int_0^T \int_{\mathbb{R}^2} \rho^{p+1}\,dx\,dt \\ \le \underline{\lim}_{\epsilon\to 0}\left(\int_{\mathbb{R}^2} \rho_\epsilon^p(T,x)\,dx + (p-1)\cos\theta \int_0^T \int_{\mathbb{R}^2} \rho_\epsilon^{p+1}\,dx\,dt\right) \\ \le \underline{\lim}_{\epsilon\to 0}\int_{\mathbb{R}^2} \rho_{0,\epsilon}^p(x)\,dx = \int_{\mathbb{R}^2} \rho_0^p\,dx. \qquad (2.56)\end{aligned}$$

Then, by summing (2.55) and (2.56), we complete the proof of Corollary 1.1. ∎

3. Proof of Theorem 1.2 and the remarks. Again the first step in the proof of Theorem 1.2 is to construct the approximate solution sequence $\{(\rho_\epsilon, u_\epsilon)\}$. By Theorem A.1, we immediately have the following lemma.

LEMMA 3.1 (solution of (1.7) with smooth data). *Let $d = 1, 2$, $\rho_0 \in L^\infty(\mathbb{R}^d)$. Then (1.7) has a global smooth solution $(\rho_\epsilon, u_\epsilon)$ such that $\rho_\epsilon, \nabla u_\epsilon \in L^\infty([0,T], H^s(\mathbb{R}^d)) \cap L^2([0,T], H^{s+1}(\mathbb{R}^d))$ for any $s > \frac{d}{2} + 1, T < \infty$, and*

$$\|\rho_\epsilon(t,\cdot)\|_{L^1} \le \|\rho_0\|_{L^1}, \quad \|\rho_\epsilon(t,\cdot)\|_{L^\infty} \le \|\rho_0\|_{L^\infty} + \epsilon, \qquad (3.1)$$

$$\int_{\mathbb{R}^d} \rho_\epsilon^2(t,x)\,dx + 2\epsilon \int_0^t \int_{\mathbb{R}^d} |\nabla \rho_\epsilon|^2\,dx\,ds \le \int_{\mathbb{R}^d} \rho_0^2\,dx + \epsilon^2 t \int_{\mathbb{R}^d} |\rho_0|\,dx. \qquad (3.2)$$

Furthermore, if $\operatorname{supp}\rho_0 \subset B_r(0)$, we denote $M = (\|\rho_0\|_{L^\infty} + 1)^{\frac{1}{d}} \|\rho_0\|_{L^1}^{1-\frac{1}{d}}$; then for $r > \epsilon$ and $(t,x) \in \Omega^o =: \{(t,x) : |x| \ge r + Mt, t \ge 0\}$, there holds

$$|\rho_\epsilon(t,x)| \le \|\rho_0\|_{L^\infty} \exp[\epsilon^{-1}(r + Mt - |x|) + t\|\rho\|_{L^\infty}] \equiv Q_\epsilon(t,x). \qquad (3.3)$$

Proof. First by the third equation of (1.7), the global existence and uniqueness of solution to (1.7) is a direct consequence of Theorem A.1. Moreover, (A.2) and (A.3) imply (3.1) and (3.2), respectively. Consequently, (A.15) implies that

$$\|u_\epsilon\|_{L^\infty} \le M.$$

Next, we rewrite the first equation of (1.7) as

$$\partial_t \rho_\epsilon + u_\epsilon S_\epsilon'(\rho_\epsilon)\nabla\rho_\epsilon + S_\epsilon(\rho_\epsilon)\rho_\epsilon = \epsilon\triangle\rho_\epsilon. \qquad (3.4)$$

Let us denote $\mathcal{L} = \partial_t + u_\epsilon S_\epsilon'(\rho_\epsilon)\nabla + S_\epsilon(\rho_\epsilon) - \epsilon\triangle$. Notice by the definition of $S_\epsilon(\xi)$ that $|S_\epsilon'(\rho_\epsilon)| \le 1$ and $\|S_\epsilon(\rho_\epsilon)\|_{L^\infty} \le \|\rho_\epsilon\|_{L^\infty}$. We find

$$\mathcal{L}Q_\epsilon \ge 0 \quad \text{on} \quad \Omega^o, \qquad (3.5)$$

$$Q_\epsilon|_{|x|=r+Mt} \ge \|\rho_0\|_{L^\infty}, \quad Q_\epsilon|_{t=0,|x|\ge r} \ge 0. \qquad (3.6)$$

Hence by the maximum principle, we have

$$\rho_\epsilon(t,x) \le Q_\epsilon(t,x) \quad \text{for all } (t,x) \in \Omega^o. \qquad (3.7)$$

Similarly, we can prove

$$-\rho_\epsilon(t,x) \le Q_\epsilon(t,x) \quad \text{for all } (t,x) \in \Omega^o. \qquad (3.8)$$

Combining (3.7) with (3.8), we get (3.3). This completes the proof of the lemma. ∎

With this lemma, by (1.7), (3.1), (3.2), and [31], we have that

$$\partial_t u_\epsilon = \partial_t \nabla \triangle^{-1} \rho_\epsilon = -\nabla \triangle^{-1} \mathrm{div}(u_\epsilon S_\epsilon(\rho_\epsilon)) + \epsilon \nabla \rho_\epsilon \tag{3.9}$$

is uniformly bounded in $L^\infty([0,T], L^2(\mathbb{R}^d))$.

While again by (1.7), (3.1), and [31], we have

$$u_\epsilon \quad \text{is uniformly bounded in} \quad L^\infty([0,T], W^{1,p}(\mathbb{R}^d)) \tag{3.10}$$

for any $2 < p < \infty$. Thus by combining (3.9), (3.10), the Lions–Aubin lemma, and the similar proof of Lemma 3 in [34], we obtain that there exists a subsequence of $\{u_\epsilon\}$, which we still denote by $\{u_\epsilon\}$, and some $u \in L^\infty([0,T], W^{1,p}(\mathbb{R}^d))$ such that

$$u_\epsilon \to u \quad \text{uniformly on any compact subset of} \quad [0,T] \times \mathbb{R}^d. \tag{3.11}$$

Trivially by (3.1), there exist $\rho, m \in L^\infty([0,T] \times \mathbb{R}^d)$ such that

$$\rho_\epsilon \rightharpoonup \rho \quad \text{weakly} \quad * \quad \text{in} \quad L^\infty([0,T] \times \mathbb{R}^d), \tag{3.12}$$

$$|\rho_\epsilon| \rightharpoonup m \quad \text{weakly} \quad * \quad \text{in} \quad L^\infty([0,T] \times \mathbb{R}^d). \tag{3.13}$$

While by the definition of $S_\epsilon(\xi)$, for any compact subset K of $[0,T] \times \mathbb{R}^2$, we have

$$\|S_\epsilon(\rho_\epsilon) - |\rho_\epsilon|\|_{L^1(K)} \to 0 \quad \text{as} \quad \epsilon \to 0. \tag{3.14}$$

Thus by combining the first equation of (1.7) with (3.11)–(3.14), we have that

$$\partial_t \rho + \mathrm{div}(um) = 0 \tag{3.15}$$

holds in the sense of distributions.

Summing up the second equation of (1.7) and (3.11) and (3.12), we get

$$u = \nabla \triangle^{-1} \rho\,. \tag{3.16}$$

Thus, in order to prove that (ρ, u) is indeed a global weak solution to (1.4), we only need to prove that $m = |\rho|$. However, only in one space dimension, and $\rho_0 \in BV(\mathbb{R})$, we can prove that $d(t,x) = |\rho|(t,x)$ for almost all $(t,x) \in \mathbb{R}^+ \times \mathbb{R}$. In order to do so, let us first present the following lemma.

LEMMA 3.2. *Let $\rho_0 \in BV(\mathbb{R})$. Then*

$$\int_{\mathbb{R}} |\partial_x \rho_\epsilon(T,x)|\,dx \le e^{3\epsilon T} \int_{\mathbb{R}} |\,d\rho_0|, \tag{3.17}$$

where $\int_{\mathbb{R}} |\,d\rho_0|$ is the total variation of ρ_0.

Proof. Let g be the solution of the adjoint equation

$$\partial_t g + u_\epsilon S_\epsilon'(\rho_\epsilon)\partial_x g - (S_\epsilon(\rho_\epsilon) + \rho_\epsilon S_\epsilon'(\rho_\epsilon))g + \epsilon \partial_{xx} g = 0, \tag{3.18}$$

with the Cauchy data $g(T,\cdot) = \gamma \in C_0^\infty(\{x : |x| < R\})$, and $\|\gamma\|_{L^\infty} \le 1$. Let $\tau = T - t$, $h = e^{-3\epsilon\tau} g$; then by (3.18), we have

$$\partial_\tau h - u_\epsilon S_\epsilon'(\rho_\epsilon)\partial_x h + (S_\epsilon(\rho_\epsilon) + \rho_\epsilon S_\epsilon'(\rho_\epsilon) + 3\epsilon)h - \epsilon \partial_{xx} h = 0. \tag{3.19}$$

We assume that h reaches its minimum value at (τ_0, x_0). Then we claim that

$$\text{either} \quad h(\tau_0, x_0) \geq 0 \qquad \text{or} \quad \tau_0 = 0. \tag{3.20}$$

Otherwise, if $h(\tau_0, x_0) < 0$ and $\tau_0 > 0$, we have by the definition of (τ_0, x_0) that

$$\partial_t h(\tau_0, x_0) = 0, \quad \partial_x h(\tau_0, x_0) = 0, \quad \partial_{xx} h(\tau_0, x_0) \geq 0, \tag{3.21}$$

which implies that

$$\{\partial_\tau h - u_\epsilon S'_\epsilon(\rho_\epsilon)\partial_x h + (S_\epsilon(\rho_\epsilon) + \rho_\epsilon S'_\epsilon(\rho_\epsilon) + 3\epsilon)h - \epsilon\partial_{xx} h\}\,(t_0, x_0) < 0$$

as $S_\epsilon(\rho_\epsilon) \geq 0, \rho_\epsilon S'_\epsilon(\rho_\epsilon) + 3\epsilon > 0$. This contradicts (3.19), which proves the claim (3.20). Hence

$$h(\tau, x) \geq \min\,(0, \min(h(0, x))) = -1,$$

which implies that

$$g(t, x) \geq -e^{3\epsilon(T-t)} \quad \text{for all } (t, x) \in [0, T] \times \mathbb{R}. \tag{3.22}$$

Exactly as in the proof of (3.22), we can also prove that

$$g(t, x) \leq e^{3\epsilon(T-t)} \quad \text{for all } (t, x) \in [0, T] \times \mathbb{R}. \tag{3.23}$$

Combining (3.22) with (3.23), we get

$$\|g(t, \cdot)\|_{L^\infty} \leq e^{3\epsilon(T-t)}, \quad 0 \leq t \leq T. \tag{3.24}$$

With (3.24) and the similar proof to (3.3), we get

$$|g(t, x)| \leq \exp[\epsilon^{-1}(R + M(T-t) - |x|) + 4(T-t)\|\rho_\epsilon\|_{L^\infty}] \tag{3.25}$$

for all $(t, x) \in \{(t, x) : |x| \geq R + M(T-t), 0 \leq t \leq T\}$.

On the other hand, taking ∂_x to the first equation of (1.7), we have

$$\partial_t(\partial_x \rho_\epsilon) + \partial_x(u_\epsilon S'_\epsilon(\rho_\epsilon)\partial_x \rho_\epsilon) + (S_\epsilon(\rho_\epsilon) + \rho_\epsilon S'_\epsilon(\rho_\epsilon))\partial_x \rho_\epsilon - \epsilon\partial_{xx}\partial_x \rho_\epsilon = 0. \tag{3.26}$$

Combining (3.18), (3.25), and (3.26), we get

$$\begin{aligned} &\int_{\mathbb{R}} \partial_x \rho_\epsilon(T, x)\gamma(x)\,dx - \int_{\mathbb{R}} \partial_x \rho_{0,\epsilon} g(0, x)\,dx \\ &= \int_0^T \int_{\mathbb{R}} (\partial_t(\partial_x \rho_\epsilon)g + \partial_x \rho_\epsilon \partial_t g)\,dx\,dt = 0. \end{aligned} \tag{3.27}$$

It follows from (3.24) that

$$\begin{aligned} \left|\int_{\mathbb{R}} \partial_x \rho_\epsilon(T, x)\gamma(x)\,dx\right| &\leq e^{3\epsilon T} \int_{\mathbb{R}^2} \int_{\mathbb{R}} |\partial_x \rho_{0,\epsilon}|\,dx \\ &\leq e^{3\epsilon T} \int_{\mathbb{R}} |d\rho_0|, \end{aligned} \tag{3.28}$$

which implies (3.17). This completes the proof of the lemma. □

We now complete the proof of Theorem 1.2.

Proof of Theorem 1.2. First by (3.17), we find that $\{\rho_\epsilon(t,x)\}$ is uniformly bounded in $L^\infty([0,T],BV(\mathbb{R}))$. While by (3.1), (3.2), and the first equation in (1.7), we find that $\{\partial_t\rho_\epsilon\}$ is uniformly bounded in $L^\infty([0,T],W^{-1,\infty}(\mathbb{R}))+L^2([0,T],H^{-1}(\mathbb{R}))$. Notice by the compact embedding theorem that $BV(\mathbb{R})\hookrightarrow\hookrightarrow L^p(\mathbb{R})$ for any $p<\infty$. Thus by the Lions–Aubin lemma and a proof similar to that of Lemma 3 in [34], we find that there exists a $\rho\in L^\infty([0,T],BV(\mathbb{R}))$ for any $T<\infty$ such that

$$\rho_\epsilon\to\rho \quad \text{in} \quad C([0,T],L^p(K)) \quad \text{for all} \quad T<\infty \tag{3.29}$$

and all compact subset K of $\mathbb{R}$. In particular, this implies that $m=|\rho|$. Thus by combining (3.15) and (3.16), we prove that (ρ,u) satisfies (1.17) in the sense of distributions. Moreover, if $\operatorname{supp}\rho_0\subset B_r(0)$, (3.3) implies that

$$\sup_{(t,x)\in\Omega^o}|\rho_\epsilon|\to 0 \quad \text{as} \quad \epsilon\to 0, \tag{3.30}$$

where Ω^o is the set defined by Lemma 3.1. Thus $\operatorname{supp}\rho(t,\cdot)\subset\{x:|x|\le r+Mt\}$. This completes the proof of the theorem. □

For $d=2$, in general, we cannot prove that $m=|\rho|$. Instead, the following proposition can be proved.

PROPOSITION 3.1. *Let $\rho_0\in L^\infty(\mathbb{R}^2)$ with $\operatorname{supp}\rho_0\subset B_r(0)$, and let us use the notation $\operatorname{supp}\rho_0^+=:\{x,\rho_0(x)>0\}$ and $\operatorname{supp}\rho_0^-=:\{x,\rho_0(x)<0\}$. Then $m=|\rho|$ for almost all $(t,x)\in D=D^+\cup D^-$, where*

$$D^+=\{(t,\Phi_t^+(x))\ :\ x\in\operatorname{supp}\rho_0^-\},\quad D^-=\{(t,\Phi_t^-(x))\ :\ x\in\operatorname{supp}\rho_0^+\},$$

with $\Phi_t^\pm(x)$ being defined by

$$\begin{cases}\frac{d\Phi_t^\pm(x)}{dt}=\pm u(t,\Phi_t^\pm(x)),\\ \Phi_t^\pm(x)|_{t=0}=x.\end{cases} \tag{3.31}$$

Proof. First by multiplying $\operatorname{sign}\rho_\epsilon$ on both sides of (1.7), we find

$$\partial_t|\rho_\epsilon|+\operatorname{div}(u_\epsilon\operatorname{sign}(\rho_\epsilon)(S_\epsilon(\rho_\epsilon)-S_\epsilon(0)))\le-|\rho_\epsilon|S_\epsilon(0)+\epsilon\triangle|\rho_\epsilon|. \tag{3.32}$$

Then by (3.1), it is trivial to prove that

$$\|\operatorname{sign}(\rho_\epsilon)(S_\epsilon(\rho_\epsilon)-S_\epsilon(0))-\rho_\epsilon\|_{L^1(K)}\to 0,\quad \|S_\epsilon(0)\rho_\epsilon\|_{L^1(K)}\to 0 \tag{3.33}$$

as $\epsilon\to 0$. Hence by summing up (3.11), (3.32), (3.33), and taking $\epsilon\to 0$ in (3.32), we find

$$\partial_t m+\operatorname{div}(u\rho)\le 0. \tag{3.34}$$

Now let $w=\rho+m$. By summing up (3.15) and (3.34), we find

$$\partial_t w+\operatorname{div}(uw)\le 0. \tag{3.35}$$

Denote $w^\epsilon(t,x)=\int_{\mathbb{R}^2}j_\epsilon(y)w(t,x-y)\,dy$, and by [22, Lemma 2.3], we find that w^ϵ satisfies

$$\partial_t w^\epsilon+\operatorname{div}(uw^\epsilon)\le R_\epsilon(t,x), \tag{3.36}$$

where $R_\epsilon(t,x) = \mathrm{div}(uw^\epsilon) - j_\epsilon * \mathrm{div}(uw)$, and $R_\epsilon \to 0$ in $L^1_{\mathrm{loc}}(\mathbb{R}^+ \times \mathbb{R}^2)$. Equation (3.36) directly implies that

$$\frac{dw^\epsilon(t,\Phi_t^+(x))}{dt} \le (\rho w^\epsilon)(t,\Phi_t^+(x)) + R_\epsilon(t,\Phi_t^+(x)). \tag{3.37}$$

On the other hand, by (3.3), we find that

$$\mathrm{supp}\rho \subset \{(t,x) : |x| \le r + Mt\} =: B.$$

Thus by [21, Lemma 1] or [33], we find that (3.31) has a unique global solution such that

$$C(T)^{-1}|x_1 - x_2|^{e^{4\pi t}} \le |\Phi_t^+(x_1) - \Phi_t^+(x_2)| \le C(T)|x_1 - x_2|^{e^{-4\pi t}}, \tag{3.38}$$

and because $\mathrm{div}u = \rho$, by [10, equation (74)], we have

$$\|R_\epsilon(t,\Phi_t^+(x))\|_{L^1(K)} \le e^{\|\rho\|_{L^\infty} t}\|R_\epsilon\|_{L^1(B_T)} \to 0 \tag{3.39}$$

as $\epsilon \to 0$, where $B_T = B \cap \{(t,x) : t \le T\}$. By summing up (3.37), (3.39), and letting $\epsilon \to 0$ in (3.37), we get for almost all $x \in \mathbb{R}^2$ that there holds

$$\begin{cases} \frac{dw(t,\Phi_t^+(x))}{dt} \le (\rho w)(t,\Phi_t^+(x)), \\ w(t,\Phi_t^+(x))|_{t=0} = 0, \quad x \in \mathrm{supp}\,\rho_0^-. \end{cases} \tag{3.40}$$

Then the Gronwall inequality implies that

$$w(t,\Phi_t^+(x)) = 0, \quad x \in \mathrm{supp}\,\rho_0^-, \quad t \in \mathbb{R}^+. \tag{3.41}$$

By (3.38) and (3.41), we get

$$d(t,x) = -\rho(t,x) \quad \text{for a.e.} \quad (t,x) \in D^+. \tag{3.42}$$

Similarly, by subtracting (3.15) from (3.34) and letting $q = m - \rho$, we have

$$\partial_t q - \mathrm{div}(uq) \le 0. \tag{3.43}$$

By the proof of (3.42), we have

$$m(t,x) = \rho(t,x) \quad \text{for a.e.} \quad (t,x) \in D^-. \tag{3.44}$$

Combining (3.42) and (3.44), we complete the proof of the proposition. □

Appendix. The construction of the approximate solutions. Let $S_\epsilon(\xi) = |\xi| * j_\epsilon(\xi)$, where $j_\epsilon(\xi)$ is the standard Friedrich mollifier with $\mathrm{supp}\, j_\epsilon(\cdot) \subset B_\epsilon(0)$. In the following, we are going to prove the global existence of smooth solutions to the following equations:

$$\begin{cases} \partial_t\rho + \mathrm{div}(uS_\epsilon(\rho)) = \epsilon\triangle\rho, \quad (t,x) \in (0,\infty)\times\mathbb{R}^d, \quad d = 1,2,3, \\ u = \nabla\triangle^{-1}\rho \quad \text{for} \quad d = 2,3, \qquad u = \int_{-\infty}^x \rho\, dy \quad \text{for} \quad d = 1, \\ \rho|_{t=0} = \rho_0. \end{cases} \tag{A.1}$$

THEOREM A.1. *Let $s > d/2+1, \rho_0 \in H^s_0(\mathbb{R}^d)$. Equation* (A.1) *has a unique global solution (ρ, u) such that $\rho, \nabla u \in L^\infty([0,T], H^s(\mathbb{R}^d)) \cap L^2([0,T], H^{s+1}(\mathbb{R}^d))$ for any $T < \infty$. Furthermore,*

$$\|\rho(t,\cdot)\|_{L^1} \le \|\rho_0\|_{L^1}, \quad \|\rho(t,\cdot)\|_{L^\infty} \le \|\rho_0\|_{L^\infty} + \epsilon. \tag{A.2}$$

$$\int_{\mathbb{R}^d} \rho^2(t,x)\,dx + 2\epsilon \int_0^t \int_{\mathbb{R}^d} |\nabla\rho|^2\,dx\,ds \le \int_{\mathbb{R}^d} \rho_0^2\,dx + \epsilon^2 t \int_{\mathbb{R}^d} |\rho_0|\,dx. \tag{A.3}$$

Proof. Following the standard argument for a nonlinear parabolic equation, we can prove the local existence and uniqueness of solution (ρ, u) to (A.1) such that $\rho, \nabla u \in L^\infty([0,T], H^s(\mathbb{R}^d)) \cap L^2([0,T], H^{s+1}(\mathbb{R}^d))$ for some positive constant T. Now let T^* be the lifespan of the solution (ρ, u) to (A.1). Then for $t < T^*$, notice the classical convex inequality: $\mathrm{sign}\rho\triangle\rho \le \triangle|\rho|$. By multiplying $\mathrm{sign}\rho$ on both sides of (A.1), we find

$$\partial_t|\rho| + \mathrm{div}(u\,\mathrm{sign}\rho(S_\epsilon(\rho) - S_\epsilon(0))) \le -|\rho|S_\epsilon(0) + \epsilon\triangle|\rho|. \tag{A.4}$$

Integrating the above inequality over $\mathbb{R}^d$, we get the first inequality of (A.2) for $t < T^*$.

Next, multiplying $p\rho^{p-1}$ on both sides of the first equation of (A.1) with p an even integer, we get

$$\partial_t\rho^p + \mathrm{div}(uF_\epsilon(\rho)) = G_\epsilon(\rho) + \epsilon p\rho^{p-1}\triangle\rho, \tag{A.5}$$

with $F_\epsilon(\rho) = \int_0^\rho pS'_\epsilon(\xi)\xi^{p-1}\,d\xi$ and $G_\epsilon(\rho) = \rho F_\epsilon(\rho) - p\rho^p S_\epsilon(\rho)$.

By the definition of $S_\epsilon(\rho)$, we have

$$G_\epsilon(\rho) = \begin{cases} p\rho\int_0^\epsilon S'_\epsilon(\xi)\xi^{p-1}\,d\xi - (p-1)\rho^{p+1} - \epsilon^p\rho + p\rho^p\int_{\mathbb{R}^d}\xi j_\epsilon(\xi)\,d\xi, & \rho \ge \epsilon, \\ p\rho\int_0^\rho S'_\epsilon(\xi)\xi^{p-1}\,d\xi - p\rho^p S_\epsilon(\rho), & |\rho| \le \epsilon \\ p\rho\int_0^{-\epsilon} S'_\epsilon(\xi)\xi^{p-1}\,d\xi + (p-1)\rho^{p+1} + \epsilon^p\rho - p\rho^p\int_{\mathbb{R}^d}\xi j_\epsilon(\xi)\,d\xi, & \rho \le -\epsilon, \end{cases}$$

which together with the simple inequalities that

$$p|\rho|^p \le (p-1)|\rho|^{p+1} + |\rho|, \quad p\left|\int_0^\epsilon S'_\epsilon(\xi)\xi^{p-1}\,d\xi\right| \le \epsilon^p, \quad S_\epsilon(\rho) \ge 0$$

gives us

$$G_\epsilon(\rho) \le \epsilon^p|\rho|. \tag{A.6}$$

Combining (A.2) with (A.6), and integrating (A.5) over $\mathbb{R}^d$, we find

$$\begin{aligned} &\int_{\mathbb{R}^d} \rho^p\,dx + p(p-1)\epsilon\int_0^t\int_{\mathbb{R}^d} \rho^{p-2}|\nabla\rho|^2\,dx\,ds \\ &\le \int_{\mathbb{R}^d}\rho_0^p\,dx + \epsilon^p\int_0^t\int_{\mathbb{R}^d}|\rho|\,dx\,ds \le \int_{\mathbb{R}^d}\rho_0^p\,dx + \epsilon^p t\int_{\mathbb{R}^d}|\rho_0|\,dx. \end{aligned} \tag{A.7}$$

In particular, by taking $p = 2$ in the above, we get (A.3). Moreover, for any compact subset K of $\mathbb{R}^d$, (A.7) implies that

$$\left(\int_K \rho^p(t,x)\,dx\right)^{\frac{1}{p}} \le \|\rho(t,\cdot)\|_{L^p} \le \|\rho_0\|_{L^p} + \epsilon t^{\frac{1}{p}}\|\rho_0\|_{L^1}^{\frac{1}{p}}. \tag{A.8}$$

Letting $p \to \infty$ in (A.8), we prove the second inequality of (A.2) for $t < T^*$.

Now let $E(t,x)$ be the fundamental solution of the heat operator $(\partial_t - \epsilon\triangle)$; then

$$E(t,x) = (4\pi\epsilon t)^{-\frac{d}{2}} e^{-\frac{|x|^2}{4\epsilon t}}, \quad t > 0, \tag{A.9}$$

by (A.1), and ρ can also be written by the following form:

$$\rho(t,x) = -\int_0^t \int_{\mathbb{R}^d} E(t-s, x-y)\mathrm{div}(uS_\epsilon(\rho))(s,y)\,dy\,ds + \int_{\mathbb{R}^d} E(t,x-y)\rho_0(y)\,dy$$

for $t < T^*$. By integrating by parts in the above formula, we find

$$\rho(t,x) = -\int_0^t \int_{\mathbb{R}^d} \nabla_x E(t-s,x-y)(uS_\epsilon(\rho))(s,y)\,dy\,ds + \rho_0(t,x), \tag{A.10}$$

and consequently,

$$\begin{aligned}\partial_x\rho(t,x) = -\int_0^t \int_{\mathbb{R}^d} &\nabla_x E(t-s,x-y)(\partial_x u S_\epsilon(\rho)\\ &+ uS'_\epsilon(\rho)\partial_x\rho)(s,y)\,dy\,ds + \partial_x\rho_0(t,x).\end{aligned} \tag{A.11}$$

Thus by the Hausdorff–Young inequality and the fact that $|S'_\epsilon(\xi)| \le 1$, we get

$$\begin{aligned}\|\partial_x\rho(t,\cdot)\|_{L^1} \le C(\sup_{t>0}\|\rho(t,\cdot)\|_{L^1} + \|u\|_{L^\infty}) \int_0^t (t-s)^{-\frac{1}{2}} &(\|\partial_x u(s,\cdot)\|_{L^\infty}\\ + \|\partial_x\rho(s,\cdot)\|_{L^1})\,ds + \|\partial_x\rho_0\|_{L^1},\end{aligned} \tag{A.12}$$

$$\begin{aligned}\|\partial_x\rho(t,\cdot)\|_{L^\infty} \le C(\|\rho(t,\cdot)\|_{L^\infty} + \|u\|_{L^\infty}) \int_0^t (t-s)^{-\frac{1}{2}} &(\|\partial_x u(s,\cdot)\|_{L^\infty}\\ + \|\partial_x\rho(s,\cdot)\|_{L^\infty})\,ds + \|\partial_x\rho_0\|_{L^\infty}.\end{aligned} \tag{A.13}$$

On the other hand, for $d = 2$ and 3, by the second equation of (A.1), we have

$$\begin{aligned}|u(t,x)| &= \left|\int_{\mathbb{R}^d} \frac{x-y}{|x-y|^d}\rho(t,y)\,dy\right|\\ &\le \|\rho(t,\cdot)\|_{L^\infty}\int_{|x-y|\le R}\frac{1}{|x-y|^{d-1}}\,dy + \frac{\|\rho(t,\cdot)\|_{L^1}}{R^{d-1}}\\ &\le CR\|\rho(t,\cdot)\|_{L^\infty} + \frac{\|\rho(t,\cdot)\|_{L^1}}{R^{d-1}}.\end{aligned} \tag{A.14}$$

Taking $R = (\frac{\|\rho(t,\cdot)\|_{L^1}}{\|\rho(t,\cdot)\|_{L^\infty}})^{\frac{1}{d}}$ in (A.14), we find

$$\|u(t,\cdot)\|_{L^\infty} \le C\|\rho(t,\cdot)\|_{L^1}^{\frac{1}{d}}\|\rho(t,\cdot)\|_{L^\infty}^{1-\frac{1}{d}} \le C\left(\|\rho(t,\cdot)\|_{L^1} + \|\rho(t,\cdot)\|_{L^\infty}\right), \tag{A.15}$$

while for $d = 1$, the second equation of (A.1) directly implies that

$$\|u(t,\cdot)\|_{L^\infty} \le \|\rho(t,\cdot)\|_{L^1}. \tag{A.16}$$

Similar to the proof of (A.15) and (A.16), we have

$$\|\partial_x u(t,\cdot)\|_{L^\infty} \le C\left(\|\partial_x\rho(t,\cdot)\|_{L^1} + \|\partial_x\rho(t,\cdot)\|_{L^\infty}\right). \tag{A.17}$$

Now let us set $y(t) = \|\partial_x \rho(t,\cdot)\|_{L^1} + \|\partial_x \rho(t,\cdot)\|_{L^\infty}$. By combining (A.2) with (A.12)–(A.17), we find

$$y(t) \le C \int_0^t (t-s)^{-\frac{1}{2}} y(s)\, ds + y_0, \quad t < T^*.$$

The Gronwall inequality yields that

$$y(t) \le y_0 e^{C\sqrt{t}}, \quad t < T^*. \tag{A.18}$$

On the other hand, standard energy estimates (see [24]) show that if $T^* < \infty$, then

$$\lim_{t \to T^*} (\|\partial_x u(t,\cdot)\|_{L^\infty} + \|\partial_x \rho(t,\cdot)\|_{L^\infty}) = \infty. \tag{A.19}$$

This contradicts (A.17) and (A.18). Thus $T^* = \infty$. This completes the proof of the theorem. □

Acknowledgments. The authors would like to thank the referees for their valuable suggestions.

REFERENCES

[1] R. Alexandre and C. Villani, *On the Boltzmann equation for long-range interaction,* Comm. Pure Appl. Math., 55 (2002), pp. 30–70.
[2] S. J. Chapman, *A mean-field model of superconducting vortices in three dimensions,* SIAM J. Appl. Math., 55 (1995), pp. 1259–1274.
[3] S. J. Chapman, *A hierarchy of models for type-*II *superconductors,* SIAM Rev., 42 (2000), pp. 555–598.
[4] S. J. Chapman, J. Rubinstein, and M. Schatzman, *A mean-field model of super-conducting vortices,* European J. Appl. Math., 7 (1996), pp. 97–111.
[5] Z. Chen and Q. Du, *A non-conforming finite element methods for a mean field model of superconducting vortices,* Math. Model. Numer. Anal., 34 (2000), pp. 687–706.
[6] M. G. Crandall, H. Ishi, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations,* Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.
[7] J. M. Delort, *Existence de nappes de tourbillon en dimension deux,* J. Amer. Math. Soc., 14 (1991), pp. 553–586.
[8] R. J. DiPerna and A. J. Majda, *Oscillations and concentrations in weak solutions of the incompressible fluid equations,* Comm. Math. Phys., 108 (1987), pp. 667–689.
[9] R. J. DiPerna and A. J. Majda, *Concentrations in regularizations for 2-D incompressible flow,* Comm. Pure Appl. Math., 40 (1987), pp. 301–345.
[10] R. J. DiPerna and P. L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces,* Invent. Math., 98 (1989), pp. 511–547.
[11] A. T. Dorsey, *Vortex motion and the Hall effect in type-*II *superconductors: A time-dependent Ginzburg-Landau theory approach,* Phys. Rev. B, 46 (1992), pp. 8376–8392.
[12] Q. Du, *Convergence analysis of a numerical method for a mean field model of superconducting vortices,* SIAM J. Numer. Anal., 37 (2000), pp. 911–926.
[13] Q. Du, M. Gunzburger, and H. Lee, *Analysis and computation of a mean field model for superconductivity,* Numer. Math., 81 (1999), pp. 539–560.
[14] Q. Du, M. D. Gunzburger, and J. S. Peterson, *Analysis and approximation of the Ginzburg–Landau model of superconductivity,* SIAM Rev., 34 (1992), pp. 54–81.
[15] W. E, *Dynamics of vortices in Ginzburg-Landau theories with applications to superconductivity,* Phys. D, 77 (1994), pp. 383–404.
[16] W. E, *Dynamics of vortex liquids in Ginzburg-Landau theories with applications to superconductivity,* Phys. Rev. B, 50 (1994), pp. 1126–1135.
[17] W. E, private communication, 1998.
[18] C. Elliott and V. Styles, *Numerical analysis of a mean field model of superconductivity,* IMA J. Numer. Anal., 21 (2001), pp. 1–51.

[19] C. M. Elliott, R. Schätzle, and B. E. E. Stoth, *Viscosity solutions of a degenerate parabolic-elliptic system arising in the mean-field theory of super-conductivity,* Arch. Ration. Mech. Anal., 145 (1998), pp. 99–127.

[20] L. C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, CBMS Reg. Conf. Ser. Math. 74, AMS, Providence, RI, 1990.

[21] F. Lin and P. Zhang, *On the hydrodynamic limit of Ginzburg-Landau vortices,* Discrete Contin. Dynam. Systems, 6 (2000), pp. 121–142.

[22] P. L. Lions, *Mathematical Topics in Fluid Mechanics, Vol.* 1, *Incompressible Models*, Oxford Lecture Ser. Math. Appl. 3, Clarendon Press, Oxford, 1996.

[23] P. L. Lions, *Mathematical Topics in Fluid Mechanics, Vol.* 2, *Compressible Models*, Oxford Lecture Ser. Math. Appl. 10, Clarendon Press, Oxford, 1998.

[24] A. J. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables,* Springer-Verlag, New York, 1984.

[25] A. J. Majda, *Vorticity and the mathematical theory of incompressible fluid flow,* Comm. Pure Appl. Math., 39 (1986), pp. 187–220.

[26] A. J. Majda, *Remarks on weak solutions for vortex sheets with a distinguished sign,* Indiana Univ. Math. J., 42 (1993), pp. 921–939.

[27] A. Majda, G. Majda, and Y. Zheng, *Concentrations in the one-dimensional Vlasov-Poisson equations.* I. *Temporal development and non-unique weak solutions in the single component case,* Phys. D, 74 (1994), pp. 268–300.

[28] A. Majda, G. Majda, and Y. Zheng, *Concentrations in the one-dimensional Vlasov-Poisson equations.* II. *Screening and the necessity for measure-valued solutions in the two component case,* Phys. D, 79 (1994), pp. 41–76.

[29] G. Richardson and B. Stoth, *Ill-posedness of the mean-field model of superconducting vortices and a possible regularisation,* European J. Appl. Math., 11 (2000), pp. 137–152.

[30] R. Schatzle and V. Styles, *Analysis of a mean field model of superconducting vortices,* European J. Appl. Math., 10 (1999), pp. 319–352.

[31] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions,* Princeton University Press, Princeton, NJ, 1970.

[32] M. E. Taylor, *Pseudo-Differential Operators and Nonlinear PDE,* Birkhäuser-Verlag, Boston, 1991.

[33] V. I. Yudovich, *Nonstationary flow of an ideal incompressible liquid,* USSR Comput. Math. Phys., 3 (1963), pp. 1407–1456.

[34] P. Zhang and Y. Zheng, *Rarefactive solutions to a nonlinear variational wave equation of liquid crystals,* Comm. Partial Differential Equations, 26 (2001), pp. 381–419.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1300–1307

OPTIMAL RATE OF CONVERGENCE FOR ANISOTROPIC VANISHING VISCOSITY LIMIT OF A SCALAR BALANCE LAW*

CHARALAMBOS MAKRIDAKIS† AND BENOÎT PERTHAME‡

Abstract. An open question in numerical analysis of multidimensional scalar conservation laws discretized on nonstructured grids is the optimal rate of convergence. The main difficulty lies on a priori BV bounds which cannot be derived by opposition to the case of structured (Cartesian) grids. In this paper we consider a related question for a corresponding continuous model, namely, the vanishing viscosity method for a multidimensional scalar conservation law with a general diffusion matrix which is only bounded. Then BV estimates are not available here; nevertheless we prove the $h^{1/2}$ convergence rate. Our strategy of proof differs from the classical method of Kuznetsov. It consists in using in an accurate way the entropy dissipation due to the parabolic terms. The dissipation of the conservation law is not strong enough, and we thus consider an auxiliary parabolic problem to compensate that. Using the kinetic formulation and the related uniqueness method also helps to avoid unessential technicalities.

Key words. rate of convergence, vanishing viscosity method, kinetic formulation, scalar conservation laws

AMS subject classifications. 35B25, 35K65, 35B45, 35L60, 65M15

PII. S0036141002407995

1. Introduction.

We consider the entropy solution

$$u \in C(\mathbb{R}^+; L^1(\mathbb{R}^d)) \cap L^\infty(\mathbb{R}^+; BV(\mathbb{R}^d))$$

to a multidimensional scalar conservation law

$$\begin{aligned} &\frac{\partial}{\partial t} u(t,x) + \operatorname{div} A(u) = 0, \qquad t > 0,\ x \in \mathbb{R}^d, \\ &\frac{\partial}{\partial t} S(u(t,x)) + \operatorname{div} \eta^S(u) \le 0 \qquad \text{for all } S \text{ convex}, \\ &u(t=0,x) = u^0(x) \in BV \cap L^\infty(\mathbb{R}^d), \end{aligned} \tag{1.1}$$

with the notation $\eta^S(u) = \int_0^u S'(\cdot) a(\cdot)$ and $a = A' : \mathbb{R} \to \mathbb{R}^d$.

A classical open question in the numerical analysis of this equation discretized on nonstructured grids is the optimal rate of convergence. Indeed, in such situations BV bounds on the numerical approximation are not available, and thus Kuznetsov's classical approach [13] does not apply and only a reduced convergence rate in $h^{1/4}$ can be established (see [5], [19] and also [7] , [11], [10], [15], [9]). This multidimensional situation is in opposition with the one-dimensional case, where such BV bounds are derived [18] and optimal rate of convergence $h^{1/2}$ follows. The result of Sanders [18] can be generalized in more than one dimension only for Cartesian grids. Recently, Cockburn and Gremaud [6], as a means of proving the optimal rate, proposed a variant

*Received by the editors May 21, 2002; accepted for publication (in revised form) October 24, 2002; published electronically May 12, 2003.
http://www.siam.org/journals/sima/34-6/40799.html
†Department of Applied Mathematics, University of Crete, 71409 Heraklion, Crete, Greece, and Institute of Applied and Computational Mathematics, FORTH, 71110 Heraklion, Crete, Greece (makr@math.uoc.gr).
‡Département de Mathématiques et Applications, UMR 8553, Ecole Normale Supérieure, 45, rue d'Ulm, 75230 Paris Cedex 05, France (perthame@dma.ens.fr).

of Kuznetsov's approach aiming to show the expected rates by bypassing the stability estimates of the approximate problem. This approach was, however, restricted to strong conditions on the mesh and the discrete fluxes.

It is usual to relate numerical methods to the vanishing viscosity method (below we always use the convention of summation upon the repeated index),

$$\begin{aligned} &\frac{\partial}{\partial t}v(t,x)+\mathrm{div}A(v)=\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}v\right), \qquad t>0,\ x\in\mathbb{R}^d,\\ &v(t=0,x)=v^0(x)\in L^1\cap L^\infty(\mathbb{R}^d), \end{aligned} \tag{1.2}$$

where the anisotropic matrix a_{ij} *reflects the unstructured character of the grid*, and thus it is only natural to assume that for some constant $K>0$,

$$a_{ij} \text{ is a positive definite symmetric matrix,} \quad \| a_{ij} \|_{L^\infty(\mathbb{R}^d)} = K. \tag{1.3}$$

Then the same difficulty appears that the standard method for error estimates does not apply.

Indeed, we recall that, as stated in a compact form in [1], Kuznetsov's result requires us to control entropies in a weak form. Namely, error terms E^S in the hyperbolic entropy inequalities, for convex S,

$$\frac{\partial}{\partial t}S(v)+\mathrm{div}\eta^S(v)\le \mathrm{div}E^S(t,x), \tag{1.4}$$

imply error estimates

$$\|u(t)-v(t)\|_{L^1(\mathbb{R}^d)}\le \|u^0-v^0\|_{L^1(\mathbb{R}^d)}+C(t)\big(\|u^0\|_{TV(\mathbb{R}^d)}\big)^{1/2}|||E|||^{1/2}, \tag{1.5}$$

with

$$|||E|||=\int_0^t\int_{\mathbb{R}^d}\sup_{|S'|\le 1,S''\ge 0}|E^S(s,x)|dx\ ds.$$

For the vanishing viscosity method (1.2), we have, for S convex,

$$\frac{\partial}{\partial t}S(v)+\mathrm{div}\eta^S(v)\le \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}S(v)\right). \tag{1.6}$$

Therefore the inequality (1.5) applies with

$$E^S(t,x)=\nabla S(v(t,x))=S'(v)\nabla v,$$

and we directly deduce the standard result

$$\begin{aligned} \|u(t)-v(t)\|_{L^1(\mathbb{R}^d)}\le{}& \|u^0-v^0\|_{L^1(\mathbb{R}^d)}\\ &+C(t)\big(\|u^0\|_{TV(\mathbb{R}^d)}\ \|v\|_{L^\infty((0,t);TV(\mathbb{R}^d))}\big)^{1/2}\big(\|a_{ij}\|_{L^\infty(\mathbb{R}^d)}\big)^{1/2}. \end{aligned}$$

With only the L^∞ assumption (1.3), we do not have a priori BV bound for the function v (except in one dimension). Therefore the general estimate (1.5) does not apply here.

The present paper develops new ideas to prove the following.

THEOREM 1.1. *For a smooth matrix a_{ij} satisfying* (1.3) *and the smooth bounded solution $v \in C(\mathbb{R}^+; L^1(\mathbb{R}^d))$ to* (1.2)*, we have*

$$\begin{aligned} \|u(t) - v(t)\|_{L^1(\mathbb{R}^d)} \le \|u^0 - v^0\|_{L^1(\mathbb{R}^d)} \\ + C(d)\|u^0\|_{TV(\mathbb{R}^d)} \left(t \,\|\, a_{ij} \,\|_{L^\infty(\mathbb{R}^d)}\right)^{1/2}. \end{aligned} \tag{1.7}$$

One of the ingredients of the proof relies on the precise entropy equality for (1.2), namely,

$$\frac{\partial}{\partial t} S(v) + \mathrm{div}\eta^S(v) = \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}S(v)\right) - S''(v)a_{ij}(x)\frac{\partial v}{\partial x_i}\frac{\partial v}{\partial x_j}. \tag{1.8}$$

In particular we have included the precise parabolic entropy dissipation term $S''(v)a_{ij}(x)\frac{\partial v}{\partial x_i}\frac{\partial v}{\partial x_j}$, which is essential in our analysis. This term has already been used in the proof of uniqueness for various hyperbolic/parabolic problems with the anisotropic nonlinear diffusions [4], and also by Chen and DiBenedetto [3] (but it can be recovered from a weaker entropy inequality for isotropic diffusions [2], [8]). Another idea developed here is that this entropy dissipation is not enough and a direct comparison with the hyperbolic solution u is not possible. In order to obtain such entropy dissipation we can only compare v with a solution to a parabolic equation with a constant diffusion term.

The proof also covers the case

$$\frac{\partial}{\partial t} v(t,x) + \mathrm{div}A(v) = \frac{\partial}{\partial x_i}\left(a_{ij}(x,v)\frac{\partial v}{\partial x_j}\right), \qquad t > 0,\ x \in \mathbb{R}^d,$$

under appropriate smoothness assumptions on v and, provided that matrix a_{ij} still is bounded, (1.3). Note that estimates of the viscosity approximation

$$\frac{\partial}{\partial t} v(t,x) + \mathrm{div}A(v) = \frac{\partial}{\partial x_i}\left(B_{ij}(v)\frac{\partial v}{\partial x_j}\right), \qquad t > 0,\ x \in \mathbb{R}^d,$$

without using the TV stability of v, were first proved in [6] in one dimension and extended to many dimensions in [1]. These proofs do not cover our case (1.3) since when applied to (1.2) they require a_{ij} to be differentiable.

Remark 1.1. The method does not use the smoothness of the function v, neither the positivity nor smoothness of the matrix a_{ij}, and it could be extended to a purely $C(\mathbb{R}^+; L^1(\mathbb{R}^d))$ setting using kinetic solutions along the lines of [17]. We have chosen, for simplicity, to use this framework on v in order to avoid unessential technicalities.

2. Proof of Theorem 1.1.

2.1. More entropy dissipation. In fact we are going to prove a variant of Theorem 1.1, comparing v with the solution $w \in C(\mathbb{R}^+; L^1(\mathbb{R}^d)) \cap L^\infty(\mathbb{R}^+; BV(\mathbb{R}^d))$ to the parabolic equation (recall the definition of K in (1.3))

$$\begin{aligned} \frac{\partial}{\partial t} w + \mathrm{div}A(w) = K\Delta w, \\ w(t=0,x) = u^0(x). \end{aligned} \tag{2.1}$$

THEOREM 2.1. *With the assumptions of Theorem* 1.1*, we have*

$$\begin{aligned} \|w(t) - v(t)\|_{L^1(\mathbb{R}^d)} \le \|u^0 - v^0\|_{L^1(\mathbb{R}^d)} \\ + C(d)\ \|u^0\|_{TV(\mathbb{R}^d)}\ (K\,t)^{1/2}. \end{aligned} \tag{2.2}$$

Theorem 1.1 follows directly from this because we can apply (1.5) to compare u and w. Since we have, for all $t \geq 0$,

$$\|w(t)\|_{TV(\mathbb{R}^d)} \leq \|u^0\|_{TV(\mathbb{R}^d)},$$

we indeed deduce from (1.5) that

$$\|w(t) - u(t)\|_{L^1(\mathbb{R}^d)} \leq C(t)\|u^0\|_{TV(\mathbb{R}^d)}\ K^{1/2}.$$

The proof is therefore reduced to proving Theorem 2.1. This will be shown in what follows; a main point here is the fact that (2.1) contains more entropy dissipation than (1.1).

2.2. Kinetic formulations. We use the kinetic framework [14], [16], [17], which simplifies very much uniqueness arguments compared to the initial Kruzhkov approach [12]. This needs some notation. We define after [14] the "equilibrium" function of density w by $\boldsymbol{\chi}(t,x,\xi) := \boldsymbol{\chi}(\xi; w(t,x))$ by

$$\boldsymbol{\chi}(\xi;w) = \begin{cases} +1 & \text{for } \ 0 < \xi < w(t,x), \\ -1 & \text{for } \ w(t,x) < \xi < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

The theory of kinetic formulations states that (2.1) is equivalent to writing the kinetic equation on $\boldsymbol{\chi}$,

$$\partial_t \boldsymbol{\chi} + a(\xi)\cdot\nabla_x \boldsymbol{\chi} = K\ \Delta \boldsymbol{\chi} + \frac{\partial}{\partial_\xi} m(t,x,\xi), \tag{2.4}$$

for some nonnegative bounded measure m given by

$$m(t,x,\xi) = K\ \delta(\xi - w(t,x))\ |\nabla w|^2. \tag{2.5}$$

The derivation of this equation from (2.1) shows that the measure m expresses the entropy dissipation. Indeed, after multiplying (2.4) by $S'(\xi)$ and ξ integration, we obtain

$$\frac{\partial}{\partial t} S(w) + \mathrm{div}\eta^S(w) = K\Delta S(w) - S''(w)\ K\ |\nabla w|^2, \tag{2.6}$$

which is the entropy equality for (2.1). Indeed, the function $\boldsymbol{\chi}$ is chosen because it provides the equalities

$$S(w) = \int_{\mathbb{R}} S'(\xi)\,\boldsymbol{\chi}(t,x,\xi)\,d\xi, \quad \eta^S(w) = \int_{\mathbb{R}} S'(\xi)\,a(\xi)\,\boldsymbol{\chi}(t,x,\xi)\,d\xi.$$

Similarly, we can perform the same construction for the function v and define, still using the notation in (2.3), $\bar{\boldsymbol{\chi}}(t,x,\xi) := \boldsymbol{\chi}(\xi; v(t,x))$. It solves

$$\partial_t \bar{\boldsymbol{\chi}} + a(\xi)\cdot\nabla_x \bar{\boldsymbol{\chi}} = \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}\bar{\boldsymbol{\chi}}\right) + \frac{\partial}{\partial_\xi}\bar{m}(t,x,\xi), \tag{2.7}$$

$$\bar{m}(t,x,\xi) = \delta(\xi - v(t,x))\ a_{ij}\frac{\partial}{\partial x_i} v \frac{\partial}{\partial x_j} v. \tag{2.8}$$

2.3. Regularization. We shall need more regularity than is available on the function $\boldsymbol{\chi}(\xi; w(t,x))$. We set $\varepsilon = (\varepsilon_1, \varepsilon_2)$, ε_1 for the forward time regularization and ε_2 for the space regularization, and we define

$$\varphi_\varepsilon(t,x) = \frac{1}{\varepsilon_1}\varphi_1\left(\frac{t}{\varepsilon_1}\right)\ \frac{1}{\varepsilon_2^d}\varphi_2\left(\frac{x}{\varepsilon_2}\right),$$

where $\varphi_j \geq 0, j = 1,2$, denote the normalized regularizing kernels with $\int \varphi_j = 1$, $\mathrm{supp}(\varphi_1) \subset (-1,0)$ in order to allow the time regularization. Next we set

$$\boldsymbol{\chi}_\varepsilon(t,x,\xi) = \boldsymbol{\chi}(\xi; w(t,x)) \star_{(t,x)} \varphi_\varepsilon. \tag{2.9}$$

The regularity of the kinetic formulation leads to an equation on $\boldsymbol{\chi}_\varepsilon$,

$$\partial_t \boldsymbol{\chi}_\varepsilon + a(\xi)\cdot\nabla_x \boldsymbol{\chi}_\varepsilon = K\Delta\boldsymbol{\chi}_\varepsilon + \frac{\partial}{\partial_\xi} m_\varepsilon(t,x,\xi), \tag{2.10}$$

$$m_\varepsilon(t,x,\xi) = m(t,x,\xi) \star_{(t,x)} \varphi_\varepsilon. \tag{2.11}$$

2.4. Decay functional. Following [16], we introduce the decay functional

$$Q_\varepsilon(t) = \int_{\mathbb{R}\times\mathbb{R}^d} [|\boldsymbol{\chi}_\varepsilon(t,x,\xi)| + |\bar{\boldsymbol{\chi}}(t,x,\xi)| - 2\boldsymbol{\chi}_\varepsilon(t,x,\xi)\ \bar{\boldsymbol{\chi}}(t,x,\xi)]\, d\xi\, dx \geq 0. \tag{2.12}$$

Since $|\boldsymbol{\chi}_\varepsilon| = sgn(\xi)\boldsymbol{\chi}_\varepsilon$, and using the L^1 assumption, which allows us to integrate by parts, we have

$$\begin{aligned}
\frac{d}{dt}Q_\varepsilon(t) &= -2\int_{\mathbb{R}\times\mathbb{R}^d} [m_\varepsilon(t,x,\xi=0) + \bar{m}(t,x,\xi=0)]\, d\xi\, dx \\
&+2\int_{\mathbb{R}^d} a_{ij}\frac{\partial}{\partial x_i}\bar{\boldsymbol{\chi}}\ \frac{\partial}{\partial x_j}\boldsymbol{\chi}_\varepsilon\ dx + 2\int_{\mathbb{R}\times\mathbb{R}^d} \bar{m}(t,x,\xi)\ \frac{\partial}{\partial_\xi}\boldsymbol{\chi}_\varepsilon\, d\xi\, dx \\
&-2\int_{\mathbb{R}\times\mathbb{R}^d} K\bar{\boldsymbol{\chi}}\ \Delta\boldsymbol{\chi}_\varepsilon\, d\xi\, dx + 2\int_{\mathbb{R}\times\mathbb{R}^d} m_\varepsilon(t,x,\xi)\ \frac{\partial}{\partial_\xi}\bar{\boldsymbol{\chi}}\, d\xi\, dx \\
&= -2\int_{\mathbb{R}^+\times\mathbb{R}^{2d}} \bar{m}(t,x,\xi=w(s,y))\ \varphi_\varepsilon(t-s,x-y)dsdydx - 2\int_{\mathbb{R}^d} m_\varepsilon(t,x,\xi=v(t,x))\ dx \\
&+2\int_{\mathbb{R}\times\mathbb{R}^d} a_{ij}\frac{\partial}{\partial x_i}\bar{\boldsymbol{\chi}}\ \frac{\partial}{\partial x_j}\boldsymbol{\chi}_\varepsilon\, d\xi\, dx - 2\int_{\mathbb{R}\times\mathbb{R}^d} K\bar{\boldsymbol{\chi}}\ \Delta\boldsymbol{\chi}_\varepsilon\, d\xi\, dx.
\end{aligned}$$

We refer to [16], [17] for justification of the significance of all these terms. Here the two negative terms containing m are favorable to proving the decay of Q_ε, and the two other terms have to be controlled, which we do now.

We begin with the worse, containing a_{ij}, which is treated in an original way here.

$$\int_{\mathbb{R}\times\mathbb{R}^d} a_{ij}\frac{\partial}{\partial x_i}\bar{\boldsymbol{\chi}}\ \frac{\partial}{\partial x_j}\boldsymbol{\chi}_\varepsilon\, d\xi\, dx$$

$$= \int_{\mathbb{R}^+\times\mathbb{R}^{2d+1}} \delta(\xi - v(t,x))\delta(\xi - w(s,y))\ a_{ij}(x)\frac{\partial}{\partial x_i}v(t,x)\ \frac{\partial}{\partial x_j}w(s,y)\ \varphi_\varepsilon(t-s,x-y)$$

$$\le \frac{1}{2}\int_{\mathbb{R}^+\times\mathbb{R}^{2d+1}} \delta(\xi - v(t,x))\delta(\xi - w(s,y))\ a_{ij}(x)\left[\frac{\partial}{\partial x_i}v(t,x)\ \frac{\partial}{\partial x_j}v(t,x)\right.$$

$$\left. + \frac{\partial}{\partial x_i}w(s,y)\ \frac{\partial}{\partial x_j}w(s,y)\right]\ \varphi_\varepsilon(t-s,x-y)\, d\xi\, dx\, dy\, ds$$

$$\le \frac{1}{2}\int_{\mathbb{R}^+\times\mathbb{R}^{2d}} \bar{m}(t,x,\xi = w(s,y))\ \varphi_\varepsilon(t-s,x-y) + \frac{1}{2}\int_{\mathbb{R}^d} m_\varepsilon(t,x,\xi = v(t,x))\ dx,$$

where we have used the definitions of m_ε and $\bar{m}$ and the bound in (1.3). Hence we conclude that

$$\frac{d}{dt}Q_\varepsilon(t) \le -2\int_{\mathbb{R}\times\mathbb{R}^d} K\bar{\boldsymbol{\chi}}\ \Delta\boldsymbol{\chi}_\varepsilon\, d\xi\, dx \le 2K\int_{\mathbb{R}\times\mathbb{R}^d} |\Delta\boldsymbol{\chi}_\varepsilon|\, d\xi\, dx. \tag{2.13}$$

To proceed further, we upper bound the right-hand side of (2.13) by

$$|\Delta\boldsymbol{\chi}_\varepsilon| = \left|\int_{\mathbb{R}^+\times\mathbb{R}^d} \Delta\boldsymbol{\chi}(s,y,\xi)\varphi_\varepsilon(t-s,x-y)ds\, dy\right|$$

$$= \left|\int_{\mathbb{R}^+\times\mathbb{R}^d} \nabla\boldsymbol{\chi}(s,y,\xi)\cdot\nabla\varphi_\varepsilon(t-s,x-y)ds\, dy\right|$$

$$= \left|\int_{\mathbb{R}^+\times\mathbb{R}^d} \delta(\xi - w(s,y))\,\nabla w(s,y)\cdot\nabla\varphi_\varepsilon(t-s,x-y)ds\, dy\right|,$$

and we conclude that

$$\int_{\mathbb{R}\times\mathbb{R}^d} |\Delta\boldsymbol{\chi}_\varepsilon|\, d\xi\, dx \le \frac{C}{\varepsilon_2}\|u^0\|_{TV(\mathbb{R}^d)},$$

$$\frac{d}{dt}Q_\varepsilon(t) \le \frac{C\,K}{\varepsilon_2}\,\|u^0\|_{TV(\mathbb{R}^d)}. \tag{2.14}$$

2.5. Conclusion of the proof. We can now conclude the proof. We deduce from (2.14) that

$$Q_\varepsilon(t) \le Q_\varepsilon(0) + \frac{C\,K\,t}{\varepsilon_2}\,\|u^0\|_{TV(\mathbb{R}^d)}.$$

On the other hand, we can upper bound the initial error by

$$Q_\varepsilon(0) = \int_{\mathbb{R}^{2d}} [|w(s,y)| + |v^0(x)| - 2\min(|w(s,y)|, |v^0(x)|)_{sgn(w(s,y)v^0(x))\ge 0}]$$

$$\varphi_\varepsilon(-s,x-y)\, dx\ dy\ ds$$

$$= \int_{\mathbb{R}^{2d}} |u^0(s,y) - v^0(x)|\ \varphi_\varepsilon(-s,x-y)\, dx\ dy\ ds.$$

At this level we may pass to limit as ε_1 vanishes, and we find (with the obvious modification on the definition of Q_{ε_2})

$$Q_{\varepsilon_2}(t) \le \|u^0 - v^0\|_{L^1(\mathbb{R}^d)} + C\varepsilon_2 \|u^0\|_{TV(\mathbb{R}^d)} + \frac{C\,K\,t}{\varepsilon_2}\,\|u^0\|_{TV(\mathbb{R}^d)}. \tag{2.15}$$

Finally, following the above lines, we lower bound $Q_{\varepsilon_2}(t)$ by

$$Q_{\varepsilon_2}(t) = \int_{\mathbb{R}^{2d}} |w(t,y) - v(t,x)| \varphi_{\varepsilon_2}(x-y)\,dx\,dy$$

$$\ge \|w(t) - v(t)\|_{L^1(\mathbb{R}^d)} - C\varepsilon_2 \|u^0\|_{TV(\mathbb{R}^d)}.$$

Together with (2.15) we find

$$\|w(t) - v(t)\|_{L^1(\mathbb{R}^d)} \le \|u^0 - v^0\|_{L^1(\mathbb{R}^d)} + C\varepsilon_2 \|u^0\|_{TV(\mathbb{R}^d)} + \frac{C\,K\,t}{\varepsilon_2}\,\|u^0\|_{TV(\mathbb{R}^d)},$$

and optimizing the parameter ε_2, we conclude the proof of Theorem 2.1.

REFERENCES

[1] F. Bouchut and B. Perthame, *Kruzhkov's estimates for scalar conservation laws revisited*, Trans. Amer. Math. Soc., 350 (1998), pp. 2847–2870.

[2] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[3] G.-Q. Chen and E. DiBenedetto, *Stability of entropy solutions to the Cauchy problem for a class of nonlinear hyperbolic-parabolic equations*, SIAM J. Math. Anal., 33 (2001), pp. 751–762.

[4] G.-Q. Chen and B. Perthame, *Kinetic formulation and well-posedness for kinetic solutions to degenerate parabolic-hyperbolic equations*, Ann. Inst. H. Poincaré, to appear.

[5] B. Cockburn, F. Coquel, and P.G. LeFloch, *Convergence of the finite volume method for multidimensional conservation laws*, SIAM J. Numer. Anal., 32 (1995), pp. 687–705.

[6] B. Cockburn and P.-A. Gremaud, *A priori estimates for numerical methods for scalar conservation laws. Part* I: *The general approach*, Math. Comp., 65 (1996), pp. 533–573.

[7] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handb. Numer. Anal. 7, P.G. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 2001, pp. 713–1020.

[8] R. Eymard, T. Gallouët, R. Herbin, and A. Michel, *Convergence of a finite volume scheme for nonlinear degenerate parabolic equations*, Numer. Math., 92 (2002), pp. 41–82.

[9] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. 118, Springer-Verlag, New York, 1996.

[10] C. Hillairet, *Finite volume schemes for a nonlinear hyperbolic equation. Convergence towards the entropy solution and error estimate*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 129–156.

[11] D. Kröner, *Numerical schemes for conservation laws*, Wiley and Teubner, Chichester, Stuttgart, 1997.

[12] S. Kruzhkov, *First order quasilinear equations with several space variables*, Math. USSR Sb., 10 (1970), pp. 217–273.

[13] N.N. Kuznetsov, *Accuracy of some approximate methods for computing the weak solutions of a first order quasilinear equation*, USSR Comp. Math. Math. Phys., 16 (1976), pp. 105–119.

[14] P.L. Lions, B. Perthame, and E. Tadmor, *A kinetic formulation of multidimensional scalar conservation laws and related questions*, J. Amer. Math. Soc., 7 (1994), pp. 169–191.

[15] M. Ohlberger, *A posteriori error estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 355–387.

[16] B. Perthame, *Uniqueness and error estimates in first order quasilinear conservation laws via the kinetic entropy defect measure*, J. Math. Pure Appl., 77 (1998), pp. 1055–1064.

[17] B. Perthame, *Kinetic Formulations of Conservation Laws*, Oxford Series in Mathematics and Its Applications, Oxford University Press, London, 2002.
[18] R. Sanders, *On convergence of monotone finite difference schemes with variable spatial differencing*, Math. Comp., 40 (1983), pp. 499–518.
[19] J.-P. Vila, *Convergence and error estimates in finite volume schemes for general multidimensional scalar conservation laws*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 267–295.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1308–1317

ASYMPTOTIC DECAY TOWARD THE RAREFACTION WAVES OF SOLUTIONS FOR VISCOUS CONSERVATION LAWS IN A ONE-DIMENSIONAL HALF SPACE*

TOHRU NAKAMURA†

Abstract. This paper is concerned with convergence rates toward the rarefaction waves of the solutions for scalar viscous conservation laws in a half space. We show that the convergence rate is $(1+t)^{-1/4}\log(2+t)$ in L^2-norm if the initial perturbation from the corresponding rarefaction waves is located in $H^1 \cap L^1$. This rate is equal to the well-known rate obtained for viscous conservation laws in the whole space. The proof is given by the combination of the standard L^2-energy method and L^1-estimate.

Key words. rarefaction wave, decay estimate, conservation laws, half space

AMS subject classifications. 35L65, 35L60

PII. S003614100240693X

1. Introduction.

We consider the initial-boundary value problem for scalar viscous conservation laws in the one-dimensional half space $\mathbb{R}_+ := (0,\infty)$:

$$\begin{cases} u_t + f(u)_x = u_{xx}, \quad x \in \mathbb{R}_+,\ t > 0, \\ u(0,t) = u_-, \quad t > 0, \\ u(x,0) = u_0(x) = \begin{cases} u_-, & x = 0, \\ \to u_+, & x \to \infty, \end{cases} \end{cases} \tag{1.1}$$

where f is a smooth function and $u_\pm$ are constants. We assume that f is strictly convex, i.e., for a certain positive constant α,

$$f''(u) \geq \alpha > 0, \tag{1.2}$$

and that the characteristic speeds $f'(u_\pm)$ satisfy

$$0 \leq f'(u_-) < f'(u_+). \tag{1.3}$$

We have from (1.2) and (1.3) that $u_- < u_+$. Under the conditions (1.2) and (1.3), it was already shown in [9] that the solutions of (1.1) converge to the corresponding rarefaction waves as $t \to \infty$. The rarefaction wave $r(x,t)$ is given as a weak solution of the Riemann problem for the corresponding hyperbolic conservation laws on the whole space:

$$\begin{cases} r_t + f(r)_x = 0, \quad x \in \mathbb{R},\ t > -1, \\ r(x,-1) = r_0^{\mathrm{R}}(x) := \begin{cases} u_-, & x < 0, \\ u_+, & x > 0. \end{cases} \end{cases} \tag{1.4}$$

*Received by the editors May 1, 2002; accepted for publication (in revised form) August 29, 2002; published electronically May 12, 2003.
http://www.siam.org/journals/sima/34-6/40693.html
†Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo 152-8552, Japan (tooru@is.titech.ac.jp).

Note here that $r(x,t)$ is a continuous function for $t \geq 0$. $r(x,t)$ is expressed explicitly for $t > -1$ by

$$r(x,t) = \begin{cases} u_-, & x \leq f'(u_-)(t+1), \\ (f')^{-1}\big(\dfrac{x}{t+1}\big), & f'(u_-)(t+1) \leq x \leq f'(u_+)(t+1), \\ u_+, & f'(u_+)(t+1) \leq x. \end{cases}$$

In the case of a one-dimensional whole space, Il'in and Oleinik [3] studied the stability of rarefaction waves. The convergence rate toward the rarefaction waves was first investigated by Harabetian [1] and has been considered by many authors [2, 10, 11]. This problem was also considered for the multidimensional conservation laws in [4, 12].

For the half space, it is shown by Liu, Matsumura, and Nishihara [9] that the asymptotic states of the solutions of (1.1) are classified into the following three cases according to the signatures of $f'(u_\pm)$: (a) $f'(u_-) < f'(u_+) \leq 0$, (b) $f'(u_-) < 0 < f'(u_+)$, and (c) $0 \leq f'(u_-) < f'(u_+)$. In case (a), the solutions of (1.1) converge to stationary waves. In case (b), the asymptotic states are superpositions of stationary waves and rarefaction waves. And case (c) yields rarefaction waves. Recently, the large time behaviors of the solutions for multidimensional conservation laws were studied by Kawashima, Nishibata, and Nishikawa [6, 7]. Their results will be published.

The main purpose of the present paper is to obtain the convergence rate for case (c). Note that the convergence rate for case (a) was also considered in [9] and that case (b) should be considered. The main theorem of the present paper is stated as follows.

THEOREM 1.1. *Suppose that* (1.2) *and* (1.3) *hold. Let* $u_0 - u_+ \in (H^1 \cap L^1)(\mathbb{R}_+)$ *and* $u_0(0) = u_-$. *Then the initial-boundary value problem* (1.1) *has a unique global solution* $u(x,t)$. *Moreover,* $u(x,t)$ *satisfies the following estimates:*

$$|u(t) - r(t)|_2 \leq C(1+t)^{-\frac{1}{4}} \log(2+t),$$
$$|u(t) - r(t)|_\infty \leq C(1+t)^{-\frac{1}{2}} \log^3(2+t),$$

where C *is a positive constant depending only on* u_0.

Notation. L^p denotes the usual Lebesgue space with the norm $|\cdot|_p$ for $1 \leq p \leq \infty$. For $m = 0, 1, \ldots$, H^m denotes the mth order Sobolev space with the norm $\|\cdot\|_m$. $C^k(I; H^m)$ denotes the space of k-times continuously differentiable functions from the interval I into H^m. We also denote generic positive constants by c and C.

The paper is outlined as follows. In section 2, we construct the "smooth approximation $w(x,t)$" of the rarefaction wave $r(x,t)$ in the same way as in [2]. Such an approximation is necessary because $r(x,t)$ is not smooth. The difficulty of the present problem comes from the boundary effects, which result from the fact that $w(0,t) \neq u_-$. Here u_- is the boundary data in (1.1). To avoid this difficulty, we construct the "modified smooth approximation $W(x,t)$" which is given by modifying $w(x,t)$ around the boundary to satisfy $W(0,t) = u_-$. This modification enables us to obtain the L^1-estimate of the perturbation in section 4. In section 3, we get the a priori estimates of the perturbation by using the standard energy method. Finally, in section 4, we show the decay estimates of the perturbation by combining the L^1-estimate with the L^2-estimate.

2. Smooth approximation and reformulation of the problem. First, we derive the smooth approximation of the rarefaction wave $r(x,t)$ by employing the idea of Hattori and Nishihara [2]. We define $\tilde{w}(x,t)$ as a solution of the Cauchy problem

$$\begin{cases} \tilde{w}_t + \tilde{w}\tilde{w}_x = \tilde{w}_{xx}, & x \in \mathbb{R},\ t > -1, \\ \tilde{w}(x,-1) = w_0^{\mathrm{R}}(x), & x \in \mathbb{R}, \end{cases} \tag{2.1}$$

where the initial data $w_0^{\mathrm{R}}(x)$ is defined by

$$w_0^{\mathrm{R}}(x) := \begin{cases} f'(u_-), & x < 0, \\ f'(u_+), & x > 0, \end{cases}$$

for the case $f'(u_-) > 0$. When $f'(u_-) = 0$, $\tilde{w}(x,t)$ defined above does not converge to the corresponding rarefaction wave fast enough around the boundary $x = 0$. Therefore, when $f'(u_-) = 0$, we need to modify $w_0^{\mathrm{R}}(x)$ as

$$w_0^{\mathrm{R}}(x) := \begin{cases} -f'(u_+), & x < 0, \\ f'(u_+), & x > 0, \end{cases}$$

for which the solution $\tilde{w}(x,t)$ of (2.1) satisfies $\tilde{w}(0,t) = 0$. Because (2.1) is the Burgers equation, we can get the explicit formula of $\tilde{w}(x,t)$ by using the Hopf–Cole transformation. Then, using this formula, we can get the optimal estimates of $\tilde{w}(x,t)$ [2]. Successively, we define a smooth approximation $w(x,t)$ of the rarefaction wave $r(x,t)$ as

$$w(x,t) := (f')^{-1}(\tilde{w}(x,t)). \tag{2.2}$$

$w(x,t)$ is well-defined since f is strictly convex. Substituting (2.2) for (2.1), we have the equation of $w(x,t)$:

$$\begin{cases} w_t + f(w)_x = w_{xx} + \dfrac{f'''(w)}{f''(w)} w_x^2, & x \in \mathbb{R},\ t > 0, \\ w(x,0) = w_0(x) := (f')^{-1}(\tilde{w}(x,0)), & x \in \mathbb{R}. \end{cases} \tag{2.3}$$

Here we summarize the well-known results for the smooth approximation $w(x,t)$ in Lemma 2.1. This lemma is proved by the direct computations of the explicit formula of $\tilde{w}(x,t)$. For details, readers are referred to [2, 8].

LEMMA 2.1. *For $1 \le p \le \infty$ and $t \ge 0$, $w(x,t)$ satisfies the following:*

(i) $0 \le w(0,t) - u_- \le Ce^{-c(1+t)}$ *for* $f'(u_-) > 0$ *and* $w(0,t) = u_-$ *for* $f'(u_-) = 0$.

(ii) $|w_x(0,t)| \le Ce^{-c(1+t)}$, $\quad |w_{xx}(0,t)| \le Ce^{-c(1+t)}$.

(iii) $|w(t) - r(t)|_p \le C(1+t)^{-\frac{1}{2}+\frac{1}{2p}}$.

(iv) $|w_x(t)|_p \le C(1+t)^{-1+\frac{1}{p}}$, $\quad |w_{xx}(t)|_p \le C(1+t)^{-\frac{3}{2}+\frac{1}{2p}}$.

(v) $w_x(x,t) > 0$ *for* $x \in \mathbb{R}$.

If the characteristic speed satisfies $f'(u_-) > 0$, $w(x,t)$ does not satisfy the boundary condition in (1.1), i.e., $w(0,t) \neq u_-$. So we need to modify $w(x,t)$ around the boundary. Our modified smooth approximation $W(x,t)$ is defined as

$$W(x,t) := w(x,t) - \psi(x,t), \tag{2.4}$$

where

$$\psi(x,t) := (w(0,t) - u_-)e^{-x}. \tag{2.5}$$

By virtue of this modification, $W(x,t)$ satisfies the boundary condition $W(0,t)=u_-$. Note that $\psi(x,t)\equiv 0$ if $f'(u_-)=0$. Substituting (2.4) for (2.3), we get the equation of $W(x,t)$:

$$\begin{cases} W_t + f(W)_x = W_{xx} - R(x,t), & x\in\mathbb{R}_+,\ t>0,\\ W(0,t)=u_-, & t>0,\\ W(x,0)=W_0(x):=w_0(x)-\psi(x,0), & x\in\mathbb{R}_+, \end{cases} \tag{2.6}$$

where $R(x,t)$ is defined as

$$R(x,t):=-\frac{f'''(w)}{f''(w)}w_x^2+\psi_t+(f(W+\psi)-f(W))_x-\psi_{xx}. \tag{2.7}$$

By using Lemma 2.1, the direct computations give the estimates of $W(x,t)$ and $R(x,t)$, as follows.

LEMMA 2.2. *For $1\le p\le\infty$ and $t\ge 0$, $W(x,t)$ and $R(x,t)$ satisfy*

(i) $|W(t)-r(t)|_p\le C(1+t)^{-\frac{1}{2}+\frac{1}{2p}}$,

(ii) $|W_x(t)|_p\le C(1+t)^{-1+\frac{1}{p}}$, $|W_{xx}(t)|_p\le C(1+t)^{-\frac{3}{2}+\frac{1}{2p}}$,

(iii) $W_x(x,t)>0$ *for* $x\in\mathbb{R}_+$,

(iv) $|R(t)|_p\le C(1+t)^{-2+\frac{1}{p}}$.

Define the perturbation $v(x,t)$ from the modified smooth approximation $W(x,t)$ as

$$v(x,t):=u(x,t)-W(x,t).$$

Since $W(x,t)$ converges to the rarefaction wave $r(x,t)$ fast enough, it suffices to obtain the decay estimates of $v(x,t)$. From (1.1) and (2.6), we have the equation of $v(x,t)$:

$$\begin{cases} v_t + (f(W+v)-f(W))_x = v_{xx}+R(x,t), & x\in\mathbb{R}_+,\ t>0,\\ v(0,t)=0, & t>0,\\ v(x,0)=v_0(x):=u_0(x)-W_0(x), & x\in\mathbb{R}_+. \end{cases} \tag{2.8}$$

Here we state an existence result for the solution $v(x,t)$ of (2.8). To this end, we define the solution space as

$$X_M(0,T)=\Big\{v\in C^0([0,T];H^1(\mathbb{R}_+))\ \Big|\ v_x\in L^2(0,T;H^1(\mathbb{R}_+))\ \text{ and } \sup_{0\le t\le T}\|v(t)\|_1\le M\Big\}$$

for positive constants T and M. Equation (2.8) is rewritten as an integral equation

$$v(x,t)=G_t*v_0+\int_0^t G_{t-\tau}*N(v(\tau))\,d\tau,$$

where $N(v)$ and G_t* are given by

$$N(v):=-(f(W+v)-f(W))_x+R,$$

$$G_t*v:=\frac{1}{\sqrt{4\pi t}}\int_0^\infty\left(e^{-(x-y)^2/4t}-e^{-(x+y)^2/4t}\right)v(y)\,dy.$$

By making use of a standard iteration method, it is shown that (2.8) has a unique solution locally in time.

PROPOSITION 2.3 (local existence). *Suppose that $v_0\in H^1(\mathbb{R}_+)$ and $v_0(0)=0$. For any $M>0$ with $\|v_0\|_1\le M$, there exists a positive time T depending on M such that* (2.8) *has a unique solution $v\in X_{2M}(0,T)$.*

3. A priori estimate. In this section, we show the a priori estimate of $v(x,t)$. The outline of the proof is similar to [4, 12], which consider the full space problems, but we also need to pay attention to the boundary effects.

PROPOSITION 3.1 (a priori estimate). *Suppose that $v \in X_M(0,T)$ is a solution of (2.8) for some positive constants T and M. Then there exists a positive constant C independent of T such that $v(x,t)$ satisfies the estimate*

$$\|v(t)\|_1^2 + \int_0^t |\sqrt{W_x(\tau)}v(\tau)|_2^2 + \|v_x(\tau)\|_1^2\,d\tau \le C(\|v_0\|_1^2 + 1). \tag{3.1}$$

Proof. First, we obtain the L^2-estimate of the perturbation $v(x,t)$. Multiplying (2.8) by v, we have

$$\Big(\frac{1}{2}v^2\Big)_t + (f(W+v) - f(W) - f'(W)v)\cdot W_x + v_x^2 + \{B(x,t)\}_x = Rv, \tag{3.2}$$

where $B(x,t)$ are boundary terms represented as

$$B(x,t) = (f(W+v) - f(W))v - \int_W^{W+v} f(s)ds + f(W)v - vv_x.$$

Note that the integration of $\{B(x,t)\}_x$ over $\mathbb{R}_+$ is equal to 0 since $v(0,t) = 0$. The second term of (3.2) is estimated below by using (1.2) and Lemma 2.2(iii) as

$$(f(W+v) - f(W) - f'(W)v)\cdot W_x \ge \frac{\alpha}{2}W_x v^2 \ge 0. \tag{3.3}$$

Here we have used the maximum principle of the parabolic equations. Integrating the right-hand side of (3.2) over $\mathbb{R}_+ \times (0,t)$ and using the Schwarz inequality and Lemma 2.2(iv), we have

$$\begin{aligned}\int_0^t\int_0^\infty |Rv|\,dxd\tau &\le \int_0^t |R(\tau)|_2|v(\tau)|_2\,d\tau \\ &\le \int_0^t \frac{1}{2}|R(\tau)|_2 + \frac{1}{2}|R(\tau)|_2|v(\tau)|_2^2\,d\tau \\ &\le C + \frac{1}{2}\int_0^t |R(\tau)|_2|v(\tau)|_2^2\,d\tau. \end{aligned}\tag{3.4}$$

Therefore, integrating (3.2) over $\mathbb{R}_+ \times (0,t)$ and using the estimates (3.3) and (3.4) yield

$$|v(t)|_2^2 + \int_0^t |\sqrt{W_x}v|_2^2 + |v_x|_2^2\,d\tau \le |v_0|_2^2 + C + \int_0^t |R|_2|v|_2^2\,d\tau. \tag{3.5}$$

Especially, we have

$$|v(t)|_2^2 \le |v_0|_2^2 + C + \int_0^t |R|_2|v|_2^2\,d\tau. \tag{3.6}$$

Applying the Gronwall inequality to (3.6), we obtain

$$|v(t)|_2^2 \le (|v_0|_2^2 + C)\exp\Big(\int_0^t |R|_2\,d\tau\Big) \le C(|v_0|_2^2 + 1),$$

where we have used Lemma 2.2(iv). Applying the above inequality to the last term of (3.5), we get the basic energy estimate

$$|v(t)|_2^2 + \int_0^t |\sqrt{W_x}v(\tau)|_2^2 + |v_x(\tau)|_2^2\,d\tau \le C(|v_0|_2^2 + 1). \tag{3.7}$$

Next, we estimate the spatial derivative of v. Multiplying (2.8) by $-v_{xx}$, we have

$$\Big(\frac{1}{2}v_x^2\Big)_t - (f(W+v) - f(W))_x v_{xx} + v_{xx}^2 - (v_t v_x)_x = -Rv_{xx}. \tag{3.8}$$

Note that the fourth term of (3.8) disappears after integrating over $\mathbb{R}_+$. By using the maximum principle of v and boundedness of W_x, the second term of (3.8) is estimated as

$$\begin{aligned} |(f(W+v) - f(W))_x v_{xx}| &\le \frac{1}{4}v_{xx}^2 + \{f'(W+v)v_x + (f'(W+v) - f'(W))W_x\}^2 \\ &\le \frac{1}{4}v_{xx}^2 + C(W_x v^2 + v_x^2). \end{aligned} \tag{3.9}$$

The right-hand side of (3.8) is estimated by the Schwarz inequality as

$$|Rv_{xx}| \le \frac{1}{4}v_{xx}^2 + R^2. \tag{3.10}$$

Apply the inequalities (3.9) and (3.10) to (3.8) and integrate the resultant inequality over $\mathbb{R}_+ \times (0,t)$. Then we obtain the estimate of v_x:

$$\begin{aligned} |v_x(t)|_2^2 + \int_0^t |v_{xx}|_2^2\,d\tau &\le |v_{0x}|_2^2 + C\int_0^t |\sqrt{W_x}v|_2^2 + |v_x|_2^2 + |R|_2^2\,d\tau \\ &\le C(\|v_0\|_1^2 + 1). \end{aligned} \tag{3.11}$$

Here the last inequality in (3.11) is given by using (3.7). Finally, adding (3.7) to (3.11) gives the desired estimate (3.1). □

The combination of Propositions 2.3 and 3.1 proves the global existence theorem.

THEOREM 3.2 (global existence). *Suppose that $v_0 \in H^1(\mathbb{R}_+)$ and $v_0(0) = 0$. Then there exists a unique global solution $v(x,t)$ of* (2.8) *satisfying*

$$v \in C^0([0,\infty); H^1(\mathbb{R}_+)), \quad v_x \in L^2(0,\infty; H^1(\mathbb{R}_+))$$

and the estimate (3.1).

4. Decay estimate. In order to derive the decay rate, we employ the L^1-estimate of v. This method is adopted from [4, 5, 12]. To this end, we define $a_\delta(v)$ and $A_\delta(v)$ as follows:

$$a_\delta(v) := (\operatorname{sgn} * \rho_\delta)(v) = \int_{-\infty}^{\infty} \operatorname{sgn}(y)\rho_\delta(v-y)dy,$$
$$A_\delta(v) := \int_0^v a_\delta(\eta)d\eta,$$

where sgn is a usual signature function defined as

$$\operatorname{sgn}(v) := \begin{cases} -1 & \text{for } v < 0, \\ 0 & \text{for } v = 0, \\ 1 & \text{for } v > 0. \end{cases}$$

ρ_δ denotes the Friedrichs mollifier defined as

$$\rho_\delta(v) := \frac{1}{\delta}\rho\left(\frac{v}{\delta}\right),$$

where ρ is a smooth function which has a compact support and satisfies $\int_{-\infty}^{\infty}\rho(x)dx = 1$. The time global solution $v(x,t)$ obtained in Theorem 3.2 satisfies the following L^1-estimate.

PROPOSITION 4.1 (L^1-estimate). *Suppose that $v_0 \in (H^1 \cap L^1)(\mathbb{R}_+)$. Then the solution $v(x,t)$ of (2.8) satisfies the estimate*

$$|v(t)|_1 \le |v_0|_1 + C\log(1+t). \tag{4.1}$$

Proof. Multiplying $a_\delta(v)$ on (2.8), we obtain

$$A_\delta(v)_t + (f(W+v) - f(W))_x a_\delta(v) - v_{xx}a_\delta(v) = R(x,t)a_\delta(v). \tag{4.2}$$

The second and third terms on the left-hand side of (4.2) are computed as

$$\begin{aligned}(f(W+v) - f(W))_x a_\delta(v) - v_{xx}a_\delta(v) &= \left\{(f(W+v) - f(W))a_\delta(v) - \int_0^v (f(W+s) - f(W))a_\delta'(s)\,ds - v_x a_\delta(v)\right\}_x \\ &\quad + \int_0^v (f'(W+s) - f'(W))a_\delta'(s)W_x\,ds + v_x^2 a_\delta'(v).\end{aligned} \tag{4.3}$$

Note that the integration of the first term on the right-hand side of (4.3) over $\mathbb{R}_+$ is equal to 0 since $v(0,t) = 0$ and $a_\delta(0) = 0$. The second and third terms on the right-hand side of (4.3) are positive since $a_\delta' \ge 0$ and $W_x \ge 0$. The right-hand side of (4.2) is estimated by Lemma 2.2(iv) as

$$\left|\int_0^t\int_0^\infty Ra_\delta(v)\,dxd\tau\right| \le \int_0^t |R(\tau)|_1\,d\tau \le C\log(1+t). \tag{4.4}$$

Integrate (4.2) over $\mathbb{R}_+\times(0,t)$ by using the above estimates and make $\delta \to 0$ afterward. This yields the desired estimate (4.1). □

Finally, we obtain the decay estimate of v. The combination of Lemma 2.2 and the following theorem immediately proves Theorem 1.1. The following theorem is proved by the same idea as that in [4, 12].

THEOREM 4.2 (decay estimate). *Suppose that $v_0 \in (H^1 \cap L^1)(\mathbb{R}_+)$. Then the solution $v(x,t)$ of (2.8) satisfies*

$$\begin{aligned}(1+t)^{\frac{1}{2}+\varepsilon}|v(t)|_2^2 + \int_0^t (1+\tau)^{\frac{1}{2}+\varepsilon}\left\{|\sqrt{W_x}v(\tau)|_2^2 + |v_x(\tau)|_2^2\right\}d\tau \\ \le C(1+t)^\varepsilon\log^2(2+t),\end{aligned} \tag{4.5}$$

$$\begin{aligned}(1+t)^{\frac{3}{2}+\varepsilon}|v_x(t)|_2^2 + \int_0^t (1+\tau)^{\frac{3}{2}+\varepsilon}\Big\{|\sqrt{W_x}v_x(\tau)|_2^2 + |v_{xx}(\tau)|_2^2 \\ + f'(u_-)v_x(0,\tau)^2\Big\}d\tau \le C(1+t)^\varepsilon\log^{10}(2+t)\end{aligned} \tag{4.6}$$

for arbitrary constant $\varepsilon \in (0, \frac{1}{2})$.

Proof. First, we obtain the decay estimate of v. Integrating (3.2) over $\mathbb{R}_+$ and using Lemma 2.2(iv), we have

$$\frac{1}{2}\frac{d}{dt}|v(t)|_2^2 + \frac{\alpha}{2}|\sqrt{W_x}v(t)|_2^2 + |v_x(t)|_2^2 \le \int_0^\infty |Rv|\,dx \le C(1+t)^{-1}|v(t)|_\infty. \tag{4.7}$$

From the Gagliardo–Nirenberg inequality, it follows that

$$|v|_2^2 + |v|_\infty \le |v|_\infty(|v|_1 + 1) \le |v_x|_2^{\frac{2}{3}}(|v|_1+1)^{\frac{4}{3}}.$$

Using this inequality, multiplying $(1+t)^{\frac{1}{2}+\varepsilon}$ by (4.7) yields

$$\begin{aligned}
&\frac{d}{dt}\Big\{(1+t)^{\frac{1}{2}+\varepsilon}|v(t)|_2^2\Big\} + (1+t)^{\frac{1}{2}+\varepsilon}\Big\{|\sqrt{W_x}v(t)|_2^2 + |v_x(t)|_2^2\Big\} \\
&\qquad \le C(1+t)^{-\frac{1}{2}+\varepsilon}(|v(t)|_2^2 + |v(t)|_\infty) \\
&\qquad \le C(1+t)^{\frac{1}{6}+\frac{1}{3}\varepsilon}|v_x(t)|_2^{\frac{2}{3}} \cdot (1+t)^{-\frac{2}{3}+\frac{2}{3}\varepsilon}(|v(t)|_1+1)^{\frac{4}{3}} \\
&\qquad \le \frac{1}{2}(1+t)^{\frac{1}{2}+\varepsilon}|v_x(t)|_2^2 + C(1+t)^{-1+\varepsilon}(|v(t)|_1+1)^2,
\end{aligned} \tag{4.8}$$

where the last inequality is obtained by using the Young inequality. Applying Proposition 4.1, (4.8) is rewritten as

$$\begin{aligned}
&\frac{d}{dt}\Big\{(1+t)^{\frac{1}{2}+\varepsilon}|v(t)|_2^2\Big\} + (1+t)^{\frac{1}{2}+\varepsilon}\Big\{|\sqrt{W_x}v(t)|_2^2 + |v_x(t)|_2^2\Big\} \\
&\qquad \le C(1+t)^{-1+\varepsilon}(|v(t)|_1+1)^2 \\
&\qquad \le C(|v_0|_1+1)^2(1+t)^{-1+\varepsilon}\log^2(2+t).
\end{aligned} \tag{4.9}$$

Integrating (4.9) over $(0,t)$, we get the estimate (4.5). In particular, we have

$$|v(t)|_2 \le C(1+t)^{-\frac{1}{4}}\log(2+t). \tag{4.10}$$

Next, we obtain the decay estimate of v_x. Integrating (3.8) over $\mathbb{R}_+$ yields

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}|v_x(t)|_2^2 - \int_0^\infty (f(W+v)-f(W))_x v_{xx}\,dx + \frac{1}{2}|v_{xx}(t)|_2^2 \le C|R(t)|_2^2 \\
&\qquad\qquad \le C(1+t)^{-3}.
\end{aligned} \tag{4.11}$$

The second term on the left-hand side of (4.11) is computed by using integration by parts as

$$\begin{aligned}
&-\int_0^\infty (f(W+v)-f(W))_x v_{xx}\,dx \\
&= -\int_0^\infty (f'(W+v)-f'(W))W_x v_{xx}\,dx - \int_0^\infty f'(W+v)\Big(\frac{1}{2}v_x^2\Big)_x\,dx \\
&= -\int_0^\infty (f'(W+v)-f'(W))W_x v_{xx}\,dx + \frac{1}{2}\int_0^\infty f''(W+v)v_x^3\,dx \\
&\qquad + \frac{1}{2}\int_0^\infty f''(W+v)W_x v_x^2\,dx + \frac{f'(u_-)}{2}v_x(0,t)^2.
\end{aligned} \tag{4.12}$$

Note that the third and fourth terms on the right-hand side of (4.12) are positive. The first term on the right-hand side of (4.12) is estimated by using the Schwarz inequality and Lemma 2.2(ii) as

$$\begin{aligned}\Big|\int_0^\infty (f'(W+v)-f'(W))W_x v_{xx}\,dx\Big| &\le C\int_0^\infty |W_x v v_{xx}|\,dx\\ &\le \frac{1}{4}|v_{xx}|_2^2 + C(1+t)^{-1}|\sqrt{W_x}v|_2^2.\end{aligned}\tag{4.13}$$

By using the Gagliardo–Nirenberg inequality and the Young inequality, the second term on the right-hand side of (4.12) is estimated as

$$\Big|\int_0^\infty f''(W+v)v_x^3\,dx\Big| \le C|v_x|_3^3 \le C|v_{xx}|_2^{\frac{7}{4}}|v|_2^{\frac{5}{4}} \le \frac{1}{4}|v_{xx}|_2^2 + C|v|_2^{10}.\tag{4.14}$$

Therefore, by using (4.12)–(4.14), (4.11) is rewritten as

$$\begin{aligned}&\frac{d}{dt}|v_x(t)|_2^2 + |\sqrt{W_x}v_x(t)|_2^2 + |v_{xx}(t)|_2^2 + f'(u_-)v_x(0,t)^2\\ &\qquad\le C\left\{(1+t)^{-3} + (1+t)^{-1}|\sqrt{W_x}v(t)|_2^2 + |v(t)|_2^{10}\right\}.\end{aligned}\tag{4.15}$$

Multiply (4.15) by $(1+t)^{\frac{3}{2}+\varepsilon}$ and integrate the resultant inequality over $(0,t)$. Then by applying (4.5) and (4.10) we have that

$$\begin{aligned}&(1+t)^{\frac{3}{2}+\varepsilon}|v_x(t)|_2^2 + \int_0^t (1+\tau)^{\frac{3}{2}+\varepsilon}\left\{|\sqrt{W_x}v_x|_2^2 + |v_{xx}|_2^2 + f'(u_-)v_x(0,\tau)^2\right\}d\tau\\ &\quad\le |v_{0x}|_2^2 + C + C\int_0^t (1+\tau)^{\frac{1}{2}+\varepsilon}\left\{|\sqrt{W_x}v|_2^2 + |v_x|_2^2\right\} + (1+\tau)^{\frac{3}{2}+\varepsilon}|v|_2^{10}\,d\tau\\ &\quad\le |v_{0x}|_2^2 + C + C(1+t)^{\varepsilon}\log^2(2+t) + C\int_0^t (1+\tau)^{-1+\varepsilon}\log^{10}(2+\tau)\,d\tau\\ &\quad\le C(1+t)^{\varepsilon}\log^{10}(2+t).\end{aligned}$$

This completes the proof. □

Acknowledgments. The author would like to express his deepest gratitude to his supervisor, Professor Shinya Nishibata, for his support. He is also very grateful to Professors Shuichi Kawashima, Kenji Nishihara, and Seiji Ukai for their helpful advice.

REFERENCES

[1] E. Harabetian, *Rarefactions and large time behavior for parabolic equations and monotone schemes*, Comm. Math. Phys., 114 (1988), pp. 527–536.

[2] Y. Hattori and K. Nishihara, *A note on the stability of the rarefaction wave of the Burgers equation*, Japan J. Indust. Appl. Math., 8 (1991), pp. 85–96.

[3] A. M. Il'in and O. A. Oleinik, *Behavior of the solution of the Cauchy problem for certain quasilinear equations for unbounded increase of the time*, Amer. Math. Soc. Transl., 42 (1964), pp. 19–23.

[4] K. Ito, *Asymptotic decay toward the planar rarefaction waves of solutions for viscous conservation laws in several space dimensions*, Math. Models Methods Appl. Sci., 6 (1996), pp. 315–338.

[5] S. Kawashima and S. Nishibata, *Shock waves for a model system of the radiating gas*, SIAM J. Math. Anal., 30 (1998), pp. 95–117.

[6] S. Kawashima, S. Nishibata, and M. Nishikawa, *Asymptotic stability of stationary waves for two-dimensional viscous conservation laws in half plane*, Discrete Contin. Dynam. Systems, to appear.

[7] S. Kawashima, S. Nishibata, and M. Nishikawa, *L^p-energy method for multi-dimensional viscous conservation laws and application to the stability of planar waves*, to appear.

[8] S. Kawashima and Y. Tanaka, *Stability of rarefaction waves for a model system of a radiating gas*, to appear.

[9] T.-P. Liu, A. Matsumura, and K. Nishihara, *Behaviors of solutions for the Burgers equation with boundary corresponding to rarefaction waves*, SIAM J. Math. Anal., 29 (1998), pp. 293–308.

[10] A. Matsumura and K. Nishihara, *Asymptotics toward the rarefaction waves of the solutions of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 3 (1986), pp. 1–13.

[11] A. Matsumura and K. Nishihara, *Global stability of the rarefaction wave of a one-dimensional model system for compressible viscous gas*, Comm. Math. Phys., 144 (1992), pp. 325–335.

[12] M. Nishikawa and K. Nishihara, *Asymptotics toward the planar rarefaction wave for viscous conservation law in two space dimensions*, Trans. Amer. Math. Soc., 352 (2000), pp. 1203–1215.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1318–1330

SHARP SOBOLEV INEQUALITY OF LOGARITHMIC TYPE AND THE LIMITING REGULARITY CONDITION TO THE HARMONIC HEAT FLOW*

TAKAYOSHI OGAWA†

Dedicated to Professor Takaaki Nishida on the occasion of his sixtieth birthday.

Abstract. We show a sharp version of the Sobolev inequality of the Beale–Kato–Majda and the Kozono–Taniuchi type in Lizorkin–Triebel space. As an application of this inequality, the regularity problem under the critical condition to the gradient flow of the harmonic map into a sphere is considered in the class $L^2(0,T;BMO(\mathbb{R}^n;\mathbb{S}^m))$, where BMO is the class of functions of bounded mean oscillations.

Key words. critical Sobolev inequalities, Lizorkin–Triebel space, interpolation inequality, harmonic heat flow, regularity criterion, bounded mean oscillation

AMS subject classifications. Primary, 35K55, 58E20; Secondary, 58J35, 46E30

PII. S0036141001395868

1. Introduction. In this paper, we discuss the regularity problem of smooth solutions to the time dependent harmonic heat flow from $\mathbb{R}^n$ into a unit sphere $\mathbb{S}^m$,

$$\begin{cases} \partial_t u - \Delta u = u(\nabla u, \nabla u), & t > 0, x \in \mathbb{R}^n, \\ u(t,x) : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{S}^m, & t > 0, x \in \mathbb{R}^n, \\ u(0,x) = u_0(x), \end{cases} \tag{1.1}$$

where $u(\nabla u, \nabla u) = u_i \sum_{1 \le l,j \le n} |\nabla_l u_j|^2$ denotes the second fundamental form on the sphere. This equation is first considered by Eells and Sampson [13] for the sake of constructing the stationary harmonic map from $\mathbb{R}^n$ into a sphere. By a simple observation, the following type of energy inequality is immediately obtained:

$$\|\nabla u(t)\|_2^2 + 2\int_0^t \|\partial_t u(\tau)\|_2^2 d\tau \le \|\nabla u_0\|_2^2, \quad t \in [0,T]. \tag{1.2}$$

Based on the above energy inequality, a weak solution is constructed in the space $L^\infty(0,T;\dot{H}^1(\mathbb{R}^n;\mathbb{S}^m))$ with $\partial_t u \in L^2(0,T;L^2(\mathbb{R}^n;\mathbb{S}^m))$. By an elegant penalizing method, the existence of a weak solution on the general compact Riemannian manifold was established by Chen and Struwe [10]. On the other hand, if the initial data is smooth, it is implicitly known that a smooth solution exists in time locally by using the Bochner-type formula (see, for example, Eells and Sampson [13] and Struwe [31]). This time the local smooth solution belongs to $u \in W^{1,\infty}(\mathbb{R}^n;\mathbb{S}^m)$, and the maximal existence time is characterized by $\|\nabla u_0\|_\infty$.

*Received by the editors October 1, 2001; accepted for publication (in revised form) September 6, 2002; published electronically May 12, 2003. This work was partially supported by Grant-in-Aid for Scientific Research of JSPS 11440057.

http://www.siam.org/journals/sima/34-6/39586.html

†Faculty of Mathematics, Kyushu University 36, Fukuoka 812-8581, Japan (ogawa@math.kyushu-u.ac.jp).

The regularity of the weak solution fails in general because of the existence of a blowing-up weak solution for large initial data. The example for the map from $\mathbb{R}^n$ to a sphere was shown by Coron and Ghidaglia [12] for $n \geq 3$ and by Chang, Ding, and Ye [9] for $n = 2$. However, some smallness assumption on the initial data or integrability condition on the solution itself may be possible to give the regularity.

This situation is related to the theory of a weak solution to the incompressible fluid mechanics. For the viscous incompressible fluid governed by the Navier–Stokes equation,

$$\begin{cases} \partial_t u - \Delta u + u \cdot \nabla u + \nabla p = 0, & t > 0, x \in \mathbb{R}^n, \\ \text{div } u = 0, & t > 0, x \in \mathbb{R}^n, \\ u(0, x) = u_0(x), \end{cases} \tag{1.3}$$

it is well known that there exists a global weak solution u based on an analogous energy inequality to (1.1) due to Leray [23]:

$$\|u(t)\|_2^2 + 2 \int_0^t \|\nabla u(\tau)\|_2^2 d\tau \leq \|u_0\|_2^2. \tag{1.4}$$

Although the full regularity of the weak solution to (1.3) remains open, there is some sufficient condition for the regularity in terms of a seminorm invariant under the scaling that maintains the equations. For the Navier–Stokes case, the equation is invariant under the scaling $u_\lambda(t, x) = \lambda u(\lambda^2 t, \lambda x)$, $p_\lambda(t, x) = \lambda^2 p(\lambda^2 t, \lambda x)$ $(\lambda > 0)$. Hence a criterion by the space-time norms such as

$$\int_0^T \||\nabla|^\alpha u(t)\|_p^\theta dt < \infty, \quad \frac{2}{\theta} + \frac{n}{p} = 1 + \alpha, \quad 2 \leq \theta < \infty,$$

gives the regularity of a weak solution. This is known as the Serrin condition (Ohyama [24], Serrin [28], Giga [17], Beirão da Veiga [2]). By observing the analogous scaling $u \to u_\lambda = u(\lambda^2 t, \lambda x)$ to (1.1) that preserves the equation, it is expected that there is a regularity criterion for (1.1) under the conditions

$$\nabla u \in L^\theta(0, T; L^p(\mathbb{R}^n)), \qquad \frac{2}{\theta} + \frac{n}{p} = 1, \quad n < p \leq \infty.$$

These conditions correspond to the Serrin criterion and are enough to show the regularity of the strong solution to (1.1).

In Kozono, Ogawa, and Taniuchi [21], the above observation is extended to an even weaker regularity criterion to the harmonic heat flow (1.1) by terms of the Besov spaces: Let $\phi_j(x)$ be the Littlewood–Paley dyadic decomposition of unity. Then the homogeneous Besov space $\dot{B}^s_{p,\rho}$ is defined by

$$\dot{B}^s_{p,\rho} = \{f \in \mathcal{Z}'(\mathbb{R}^n) : \|f\|_{\dot{B}^s_{p,\rho}} < \infty\},$$

where $\|f\|_{\dot{B}^s_{p,\rho}} = \left(\sum_{j=-\infty}^{\infty} 2^{js\rho} \|\phi_j * f\|_p^\rho\right)^{1/\rho}$ and $\mathcal{Z}'(\mathbb{R}^n)$ denotes the coefficient space of $\mathcal{S}'$ by the polynomials $\mathcal{P}$. The regularity criterion in the Besov space obtained in [21] is as follows.

PROPOSITION 1.1 (Kozono, Ogawa, and Taniuchi [21]). *Let u be a smooth solution to (1.1) in $C([0,T);W^{1,\infty}(\mathbb{R}^n;\mathbb{S}^m))\cap C^1((0,T);W^{2,\infty}(\mathbb{R}^n;\mathbb{S}^m))$ with initial data $u_0\in W^{1,\infty}(\mathbb{R}^n;\mathbb{S}^m)$. Suppose that the solution u satisfies either*

(i) *for any pair of (p,θ) with $\frac{2}{\theta}+\frac{n}{p}=1$ and $n<p<\infty$ and for any $\sigma\le 2p/n$,*

$$\int_0^T \|\nabla u(\tau)\|_{\dot B^0_{p,\sigma}}^{\theta} d\tau<\infty \tag{1.5}$$

or

(ii)

$$\int_0^T \|\nabla u(\tau)\|_{\dot B^0_{\infty,2}}^{2} d\tau<\infty. \tag{1.6}$$

Then the solution can be extended after $t=T$, namely, for some $T<\tilde T$, $u\in C([0,\tilde T);W^{1,\infty}(\mathbb{R}^n;\mathbb{S}^m))\cap C^1((0,\tilde T);W^{2,\infty}(\mathbb{R}^n;\mathbb{S}^m))$. In other words, if the solution blows up at $t=T$, then

$$\int_0^T \|\nabla u(\tau)\|_{\dot B^0_{p,\sigma}}^{\theta} d\tau=\infty$$

for any pair of (p,θ) satisfying $2/\theta+n/p=1$ with $\sigma\le 2p/n$ if $p<\infty$ and $\sigma=2$ if $p=\infty$.

The analogous regularity criterion to the Navier–Stokes equations is established in the scale where the equation remains invariant under the scaling [20], [27]. Among others, the corresponding condition involving bounded mean oscillation (BMO) is considered [21] (cf. for the Euler equations [22]). More precisely, the Leray weak solution is regular up to $t=T$ under the condition

$$\int_0^T \|\mathrm{rot}\ u(t)\|_{BMO}dt<\infty.$$

Here BMO is the space of the *bounded mean oscillation* defined by

$$f\in L^1_{loc}(\mathbb{R}^n)\quad \sup_{x,R}\frac{1}{|B_R|}\int_{B_R(x)}|f(y)-\bar f_{B_R(x)}|dy<\infty,$$

where $\bar f_{B_R}$ is the average of f over $B_R(x)=\{y\in\mathbb{R}^n;|x-y|<R\}$. We see that there is a gap comparing the result for the Navier–Stokes equations with the one to the harmonic heat flow, namely, the criterion at the limiting case $p=\infty$, $\theta=2$. By the strict inclusion $\dot B^0_{\infty,2}\subsetneq BMO$ (cf. Strichartz [30], Bergh and Löfström [3]), the result for (1.1) in [21] is slightly weaker than the one for the Navier–Stokes equations in view of $\dot B^0_{\infty,2}\subsetneq BMO$. This gap appears between the two cases because of the balance between the order of the nonlinearity and the order of the interpolation inequalities. To see this, we recall the basic argument found in [1] and [22]. In both results, the critical Sobolev embedding inequality of logarithmic type plays the crucial role of the regularity criterion.

Namely, the inequality originally due to Brezis and Gallouet [4], Beale, Kato, and Majda [1], and Kozono and Taniuchi [22] is suitable only for the quadratic order of nonlinearity like in the Navier–Stokes equations (1.3).

In the result of Beale, Kato, and Majda [1] (see also Kato and Ponce [20]), they showed that, for $f \in \{W^{s,p}(\mathbb{R}^n)\}^n$ $(s > n/p + 1)$ with div $f = 0$,

$$\|\nabla f\|_\infty \le C\left\{1 + \|\nabla f\|_2 + \|\omega\|_\infty \log(e + \|f\|_{W^{s+1,p}})\right\}, \quad \omega = \operatorname{rot} f. \tag{1.7}$$

An even more improved version of the inequality due to Kozono and Taniuchi [22] states that, for $f \in \{W^{s,p}(\mathbb{R}^n)\}^n$ $(s > n/p + 1)$ with div $f = 0$, there holds

$$\|f\|_\infty \le C\left\{1 + \|f\|_{BMO} \log(e + \|f\|_{W^{s,p}})\right\}. \tag{1.8}$$

However, for the regularity problem (1.1) under the condition

$$\int_0^t \|\nabla u(t)\|_{BMO}^2 dt < \infty,$$

inequalities (1.7) and (1.8) are not sufficient.

One way to fill this gap is to improve the Sobolev inequality (1.8). We first introduce a generalized version of the critical Sobolev inequality in the Lizorkin–Triebel space that includes the above inequalities. It then turns out that the second exponent of those spaces gives an explicit dependence of the logarithmic order of higher regularity, which reflects hypotheses on the integral exponent in the time direction of those criteria. In the following section, we show a refined version of the Beale–Kato–Majda- and Kozono–Taniuchi-type inequalities and give some discussion. Then, in section 3, we show our new regularity criterion for each of the problems of (1.1). The statement reads as follows.

THEOREM 1.2 (limiting regularity criterion). *Let u be a smooth solution to (1.1) in $C([0,T); W^{1,\infty}(\mathbb{R}^n;\mathbb{S}^m)) \cap C^1((0,T); W^{2,\infty}(\mathbb{R}^n;\mathbb{S}^m))$ with initial data $u_0 \in W^{1,\infty}(\mathbb{R}^n;\mathbb{S}^m)$. Suppose that the solution u satisfies*

$$\int_0^T \|\nabla u(\tau)\|_{BMO}^2 d\tau < \infty. \tag{1.9}$$

Then the solution can be extended after $t = T$, namely, for some $T < \tilde{T}$, $u \in C([0,\tilde{T}); W^{1,\infty}(\mathbb{R}^n;\mathbb{S}^m)) \cap C^1((0,T); W^{2,\infty}(\mathbb{R}^n;\mathbb{S}^m))$. In other words, if the solution blows up at $t = T$, then

$$\int_0^T \|\nabla u(\tau)\|_{BMO}^2 d\tau = \infty$$

for any pair of (p,θ) satisfying $2/\theta + n/p = 1$ and $\sigma \le 2p/n$ and $\sigma = 2$ if $p = \infty$.

It is important to compare the results of the existence of blowing-up solutions for (1.1) to the above criterion. There are several results for constructing the finite time blow-up of the solution. Coron and Ghidaglia [12] and Chen and Ding [8] showed that there exists a finite time blowing-up solution to (1.1) for $n \ge 3$. For $n = 2$, Chang, Ding, and Ye [9] constructed a blowing-up solution from a smooth data (cf. for the regularity of the stationary harmonic maps Hélein [19], Evans [15], and Coifman et al. [11], and see also Feldman [16] for the time dependent case). The solution satisfies

$$\int_0^T \|\nabla u(t)\|_\infty^r dt = \infty \quad (r > 1),$$

where $T > 0$ is the expected blow-up time. We simply remark that, for the two dimensional case, if we make the stronger regularity assumption that

$$\int_0^T \|\Delta u(t)\|_2^2 dt < \infty, \tag{1.10}$$

then, by the embedding,

$$\int_0^T \|\nabla u(t)\|_{BMO}^2 dt < \infty,$$

and our criterion gives the regularity. Because the weak solution satisfies the energy inequality, we have

$$\int_0^T \|\partial_t u(t)\|_2^2 dt < \infty.$$

Hence, if the nonlinearity has the integrability condition

$$\int_0^T \|\nabla u(t)\|_4^4 dt < \infty,$$

then the condition (1.10) is fulfilled, and the solution has to be smooth near $t = T$. This is nothing but the case of the criterion from the scaling invariant norm

$$\int_0^T \|\nabla u(t)\|_p^\theta dt < \infty, \quad \frac{2}{\theta} + \frac{n}{p} = 1.$$

Our criterion Theorem 1.2 is a stronger result and is outside of this kind of regularity criterion. We should also remark on the related results for the Euler equation. Chemin [7] considered the Euler equation in the Zygmund and log-Lipschitz class. His argument also includes the logarithmic type functional inequality in terms of the log-Lipschitz seminorm and Bony's para-product formula. Vishik [33] also develops this direction in the two dimensional case. Some related uniqueness result was shown by Yudovich [34] and Ogawa and Taniuchi [26] for the two dimensional unbounded vorticity solution.

Before closing this section, we introduce some notation. $\mathcal{F}f$ and $\hat{f}$ denote the Fourier transform of f. $\langle x \rangle = (1+|x|^2)^{1/2}$. We define a saturated logarithmic function $\log^+ t = \log(e + t)$. The usual Sobolev space $W^{s,p}(\mathbb{R}^n)$ is abbreviated as $W^{s,p}$ with the norm

$$\|f\|_{W^{s,p}} \equiv \|\mathcal{F}^{-1}\langle\cdot\rangle^s \hat{f}(\cdot)\|_p$$

for $1 < p < \infty$ and $s \geq 0$.

We recall the Paley–Littlewood dyadic decomposition (cf. Stein [29] and Bergh and Löfström [3]). Let $\phi_j(x)$ be the inverse Fourier transform of the jth component of the dyadic decomposition, i.e., $\sum_{j=-\infty}^{\infty} \hat{\phi}(2^{-j}\xi) = 1$ except $\xi = 0$, where the support

of $\hat{\phi}(\xi)$ is located on $2^{-1} < |\xi| < 2$. We denote $\psi(x) = \mathcal{F}^{-1}[\hat{\psi}(\xi)](x)$, where

$$\hat{\psi} = \begin{cases} 1, & |\xi| < 1, \\ \text{smooth}, & |\xi| < 2, \\ 0, & |\xi| > 2. \end{cases}$$

Set $\psi_j = \mathcal{F}^{-1}[\hat{\psi}(2^j\xi)](x)$. For a smooth function f, we set $\Phi_j f = \phi_j * f$ and $\Psi(x)f = \psi * f$. The homogeneous Besov space $\dot{B}^s_{p,\rho}$ is defined through the full dyadic decomposition by

$$\dot{B}^s_{p,\rho} = \{f \in \mathcal{Z}'(\mathbb{R}^n) : \|f\|_{\dot{B}^s_{p,\rho}} < \infty\},$$

where $\|f\|_{\dot{B}^s_{p,\rho}} = \left(\sum_{j=-\infty}^{\infty} 2^{js\rho}\|\phi_j * f\|_p^\rho\right)^{1/\rho}$ and $\mathcal{Z}'(\mathbb{R}^n)$ denotes the dual space of $\mathcal{Z}(\mathbb{R}^n) = \{f \in \mathcal{S}; D^\alpha \hat{f}(0) = 0 \,\forall \alpha \in \mathbb{N}^n \text{ multiindex}\}$ and can be identified by the coefficient space of $\mathcal{S}'/\mathcal{P}$ with the polynomial space $\mathcal{P}$. The homogeneous Lizorkin–Triebel space $\dot{F}^s_{p,\rho}$ is similarly defined by

$$\dot{F}^s_{p,\rho} = \{f \in \mathcal{Z}'(\mathbb{R}^n) : \|f\|_{\dot{F}^s_{p,\rho}} < \infty\},$$

where $\|f\|_{\dot{F}^s_{p,\rho}} = \|\left(\sum_{j=-\infty}^{\infty} 2^{js\rho}|\phi_j * f|^\rho\right)^{1/\rho}\|_p$ and $1 \le p < \infty$, $1 \le \rho \le \infty$ ($1 \le \rho < \infty$ if $p = \infty$). We refer to Triebel [32] for more detailed properties of those spaces.

2. Sharp version of logarithmic inequality. In this section, we give a sharp version of the logarithmic Sobolev inequality. The original type of Sobolev inequality was found by Brezis and Gallouet [4] and Brezis and Wainger [5] (see also Engler [14]). And the similar type of inequality we shall discuss here was first established by Beale, Kato, and Majda [1] and was improved by Kozono and Taniuchi [22] and Kozono, Ogawa, and Taniuchi [21]. We show the sharp version of the Kozono–Taniuchi inequality.

THEOREM 2.1 (sharp version of logarithmic inequality). (1) *For any $p, \rho, \sigma \in [1, \infty]$, $q \in [1, \infty)$, $\nu \le \sigma_1, \sigma_2$, $\nu < \rho$, and $\gamma > 0$, there exists a constant C which depends only on n, p such that, for $f \in \dot{F}^{\gamma}_{p,\sigma_1} \cap \dot{F}^{-\gamma}_{p,\sigma_2}$, we have*

$$\|f\|_{\dot{F}^0_{p,\nu}} \le C\|f\|_{\dot{F}^0_{p,\rho}} \left(1 + \left(\frac{1}{\gamma}\log^+ \frac{\|f_+\|_{\dot{F}^{\gamma}_{p,\sigma_1}} + \|f_-\|_{\dot{F}^{-\gamma}_{p,\sigma_2}}}{\|f\|_{\dot{F}^0_{p,\rho}}}\right)^{1/\nu - 1/\rho}\right), \tag{2.1}$$

where $f_+ = \sum_{j>0} \phi_j * f$ and $f_- = \sum_{j<0} \phi_j * f$.

Remark 2.1. In the theorem, the assumption $\gamma > 0$ is essential. The analogous version of the inequality (2.1) in the Besov space was proved in Ogawa and Taniuchi [25].

Proof of Theorem 2.1. To show Theorem 2.1, we recall the definition of the Lizorkin–Triebel (semi)norm. We decompose f into the following three parts: Noting that $\nu < \rho, \sigma_1, \sigma_2$, we have

(2.2)

$$
\begin{aligned}
\|f\|_{\dot{F}^0_{p,\nu}} &\le \left\| \left(\sum_{j>N} |\phi_j * f|^\nu \right)^{1/\nu} \right\|_p + \left\| \left(\sum_{|j|\le N} |\phi_j * f|^\nu \right)^{1/\nu} \right\|_p + \left\| \left(\sum_{j<-N} |\phi_j * f|^\nu \right)^{1/\nu} \right\|_p \\
&\le \left\| \left(\sum_{j>N} 2^{-j\gamma(1/\nu-1/\sigma_1)} \right)^{1/\nu-1/\sigma_1} \left(\sum_{j>N} 2^{j\gamma\sigma_1} |\phi_j * f|^{\sigma_1} \right)^{1/\sigma_1} \right\|_p \\
&\quad + (2N+1)^{1/\nu-1/\rho} \left\| \left(\sum_{|j|\le N} |\phi_j * f|^\rho \right)^{1/\rho} \right\|_p \\
&\quad + \left\| \left(\sum_{j<-N} 2^{j\gamma(1/\nu-\sigma_2)} \right)^{1/\nu-1/\sigma_2} \left(\sum_{j<-N} 2^{-j\gamma\sigma_2} |\phi_j * f|^{\sigma} \right)^{1/\sigma_2} \right\|_p \\
&\le 2^{-\gamma N} \left\{ \left\| \left(\sum_{j>N} 2^{j\gamma\sigma_1} |\phi_j * f|^{\sigma_1} \right)^{1/\sigma_1} \right\|_p + \left\| \left(\sum_{j<-N} 2^{-j\gamma\sigma_2} |\phi_j * f|^{\sigma_2} \right)^{1/\sigma_2} \right\|_p \right\} \\
&\quad + (2N+1)^{1/\nu-1/\rho} \left\| \left(\sum_{|j|\le N} |\phi_j * f|^\rho \right)^{1/\rho} \right\|_p \\
&\le 2^{-\gamma N} \left\{ \|f_+\|_{\dot{F}^{\gamma}_{p,\sigma_1}} + \|f_-\|_{\dot{F}^{-\gamma}_{p,\sigma_2}} \right\} + (2N+1)^{1/\nu-1/\rho} \|f\|_{\dot{F}^0_{p,\rho}}.
\end{aligned}
$$

Now we optimize (2.2) for each f by setting $N = 1$ if

$$\|f_+\|_{\dot{F}^{\gamma}_{p,\sigma}} + \|f_-\|_{\dot{F}^{\gamma}_{p,\sigma}} \le \|f\|_{\dot{F}^0_{p,\rho}}$$

and

$$N \simeq \left[\log_{2^\gamma} \left(\frac{\|f_+\|_{\dot{F}^{\gamma}_{p,\sigma_1}} + \|f_-\|_{\dot{F}^{-\gamma}_{p,\sigma_2}}}{\|f\|_{\dot{F}^0_{p,\rho}}} \right) \right] + 1$$

otherwise. □

Some minor modification shows that the exponents of the higher regularity for f can be chosen arbitrarily under the following form.

COROLLARY 2.2. *There exists a constant C which depends only on n, p such that, for $f \in \dot{F}^{n/p+\gamma}_{p,\sigma_1} \cap \dot{F}^{n/p-\gamma}_{p,\sigma_2}$, we have for $\gamma < \gamma'$*

$$\|f\|_{\dot{F}^0_{\infty,\nu}} \le C \|f\|_{\dot{F}^0_{\infty,\rho}} \left(1 + \left(\frac{1}{\gamma} \log^+ \frac{\|f_+\|_{\dot{F}^{n/p+\gamma'}_{p,\sigma_1}} + \|f_-\|_{\dot{F}^{n/p-\gamma'}_{p,\sigma_2}}}{\|f\|_{\dot{F}^0_{\infty,\rho}}} \right)^{1/\nu-1/\rho} \right), \tag{2.3}$$

where $f_+ = \sum_{j\ge 0} \phi_j * f$ and $f_- = \sum_{j\le 0} \phi_j * f$.

The relation between the Lizorkin–Triebel spaces and the BMO is well understood. The following result is due to Peetre and Triebel (see also Qui [6]).

PROPOSITION 2.3 (Triebel [32]). *$\dot{F}^0_{\infty,2} \simeq BMO$. Namely, there exists a constant C such that*

$$C^{-1}\|f\|_{\dot{F}^0_{\infty,2}} \le \|f\|_{BMO} \le C\|f\|_{\dot{F}^0_{\infty,2}}.$$

From (2.3) and the equivalence between $\dot{F}^0_{\infty,2} \simeq BMO$ and $\dot{F}^0_{\infty,\infty} \simeq \dot{B}^0_{\infty,\infty}$, it is explicitly shown that the difference between L^∞, BMO, and the Besov space $\dot{B}^0_{\infty,\infty}$ is as follows. This is a version of the sharp form of the Kozono–Taniuchi inequality (1.8).

COROLLARY 2.4. *We have the following: For $\gamma' > 0$,*

$$\|f\|_{BMO} \le C\left(1 + \|f\|_{\dot{B}^0_{\infty,\infty}}\left(\frac{1}{\kappa}\log^+(\|f_+\|_{\dot{F}^\kappa_{\infty,\sigma_1}} + \|f_-\|_{\dot{F}^{-\kappa}_{\infty,\sigma_2}})\right)^{1/2}\right), \tag{2.4}$$

and if $\hat{f}(0) = 0$,

$$\|f\|_\infty \le C\left(1 + \|f\|_{BMO}\left(\frac{1}{\kappa}\log^+(\|f_+\|_{\dot{F}^\kappa_{\infty,2}} + \|f_-\|_{\dot{F}^{-\kappa}_{\infty,2}})\right)^{1/2}\right). \tag{2.5}$$

In particular, if $\nabla f \in W^{1,q}(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$ for $n < q$, we have

$$\|\nabla f\|_\infty \le C(q)\left(1 + \|\nabla f\|_{BMO}\left(\log^+(\|\nabla f\|_{W^{1,q}} + \|f\|_\infty)\right)^{1/2}\right). \tag{2.6}$$

Remark 2.2. The last inequality (2.5) improves the related logarithmic inequalities (1.7) and (1.8) due to Beale, Kato, and Majda and Kozono and Taniuchi. Recalling the Brezis–Gallouet inequality,

$$\|f\|_\infty \le C(1 + \|f\|_2 + \|\nabla f\|_2(\log^+\|\Delta f\|_2)^{1/2}), \quad f \in H^2(\mathbb{R}^2),$$

one may notice that inequality (2.6) has the same order of the higher regular term despite the dimension independence, although it is substituted by the Dirichlet norm instead of the BMO seminorm.

Proof of Corollary 2.4. Noting the inequality

$$x\left(\log\left(e + \frac{y}{x}\right)\right)^{1/2} \le \begin{cases} C(1 + x(\log(e+y))^{1/2}) & \text{for } 0 < x \le 1, \\ Cx\,(\log(e+y))^{1/2} & \text{for } 1 < x, \end{cases}$$

the first inequality (2.4) is an immediate consequence of (2.1) with $\nu = 2$ and $\rho = \infty$ and Proposition 2.3. Similarly, the second inequality (2.4) follows from (2.1) with $\nu = 1$ and $\rho = 2$, observing that

$$\|f\|_\infty = \left\|\sum_{i=-\infty}^{\infty} \phi_j * f\right\|_\infty \le \|f\|_{\dot{F}^0_{\infty,1}}$$

when $\hat{f}(0) = 0$.

To obtain the last modification (2.6), we first notice

$$\lim_{M\to\infty}\left\|\nabla f - \left(\sum_{j\ge -M}\phi_j * f\right)\right\|_\infty = 0.$$

To see this, we define a smooth function $\psi(x)$ such that

$$\hat{\psi}(\xi) = \begin{cases} 1, & |\xi| \le 1/2, \\ 0, & |\xi| \ge 1, \end{cases}$$

and we set $\widehat{\psi}_j(\xi) = \widehat{\psi}(\xi/2^j)$. Then we have by the L^1-L^∞ bound of the Fourier inverse transform

$$\begin{aligned} \|\psi_{-M} * \nabla f\|_\infty &\le C_n \|\widehat{\psi}_{-M} \xi \hat{f}\|_1 \\ &\le C_n \int_{B_{2^{-M}}} |\xi \hat{f}(\xi)| d\xi \\ &\le C_n |B_{2^{-M}}|^{1/2} \left(\int_{B_{2^{-M}}} |\xi|^2 |\hat{f}(\xi)|^2 d\xi \right)^{1/2} \\ &\le C_n 2^{-Mn/2} \|\nabla f\|_2 \to 0, \end{aligned} \tag{2.7}$$

as $M \to \infty$. Hence, for sufficiently large M such that $\|\psi_{-M} * \nabla f\|_\infty \le 1$, it suffices to estimate $\sum_{j \ge -M} \phi_j * f$. We apply inequality (2.4) with small κ specified below. For small $\kappa > 0$ and $\alpha > 0$ with $\kappa < \alpha < 1 - n/q$,

$$\begin{aligned} \|\nabla f_+\|_{\dot{F}^\kappa_{\infty,2}} &= \left\| \left(\sum_{j=1}^\infty 2^{2j\kappa} |\phi_j * \nabla f|^2 \right)^{1/2} \right\|_\infty \\ &\le \left(\sum_{j=1}^\infty 2^{2j(\kappa-\alpha)} \right)^{1/2} \left\| \sup_j 2^{\alpha j} |\phi_j * \nabla f| \right\|_\infty \\ &\le C \|\nabla f\|_{\dot{B}^\alpha_{\infty,\infty}} \le C \|\nabla f\|_{\dot{B}^{\alpha+n/q}_{q,\infty}} \\ &\le C \|\nabla f\|_{\dot{W}^{1,q}}, \end{aligned} \tag{2.8}$$

where $\|\cdot\|_{\dot{W}^{1,q}}$ stands for the homogeneous Sobolev seminorm. This is possible under the condition $n < q$. On the other hand, using the L^∞ boundedness of the Hardy–Littlewood maximal function (cf. Stein [29, p. 62–63]), we have for small $0 < \kappa < 1$

$$\begin{aligned} \|\nabla f_-\|_{\dot{F}^{-\kappa}_{\infty,2}} &= \left\| \left(\sum_{j=-1}^{-\infty} 2^{-2j\kappa} |\phi_j * \nabla f|^2 \right)^{1/2} \right\|_\infty \\ &\le \left\| \left(\sum_{j=-1}^{-\infty} 2^{2j(1-\kappa)} |(\nabla \phi)_j * f|^2 \right)^{1/2} \right\|_\infty \\ &\le \left(\sum_{j=-1}^{-\infty} 2^{2j(1-\kappa)} \right)^{1/2} \left\| \sup_j |(\nabla \phi)_j * f| \right\|_\infty \\ &\le C \|M[f]\|_\infty \le C \|f\|_\infty, \end{aligned} \tag{2.9}$$

where $(\nabla\phi)_j(x) = 2^{nj} \nabla\phi(2^j x)$. From (2.8) and (2.9), we obtain the last inequality (2.6). □

3. The harmonic heat flow. In this section, we give the proof of the regularity criterion to the weak solution of the harmonic heat flow equation onto a sphere:

$$\begin{cases} \partial_t u - \Delta u = u(\nabla u, \nabla u), & t > 0, x \in \mathbb{R}^n, \\ u(t,x) : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{S}^m, & t > 0, x \in \mathbb{R}^n, \\ u(0,x) = u_0(x), \end{cases} \tag{3.1}$$

where $u(\nabla u, \nabla u) = u_i \sum_{1 \le l,j \le n} |\nabla_l u_j|^2$ denotes the second fundamental form on the sphere, and in the following we express this form as $u|\nabla u|^2$ except when it may cause confusion.

The proof is in fact a simple application to the argument in the previous section.

Proof of Theorem 1.2. We first give a proof for $n < p < \infty$. Let u be a smooth solution to (3.1) on $[0, T)$. By operating the Laplacian to the equation and then taking an L^2 inner product of the equation with $|\Delta u|^{q-2}\Delta u$, we have

$$\begin{aligned} \frac{1}{q}\frac{d}{dt}\|\Delta u(t)\|_q^q &+ \int_{\mathbb{R}^n} \nabla_k \Delta u(t) \cdot \nabla_k (|\Delta u|^{q-2}\Delta u(t))dx \\ &= \int_{\mathbb{R}^n} |\nabla u(t)|^2 \Delta u(t) \cdot |\Delta u(t)|^{q-2}\Delta u(t)dx \\ &\quad + \int_{\mathbb{R}^n} \nabla_k u(t) \cdot \nabla_k |\nabla u(t)|^2 |\Delta u(t)|^{q-2}\Delta u(t)dx \\ &\quad - \int_{\mathbb{R}^n} u(t)\nabla_k |\nabla_l u(t)|^2 \cdot \nabla_k (|\Delta u(t)|^{q-2}\Delta u(t))dx \\ &= (|\nabla u(t)|^2, |\Delta u(t)|^q) \\ &\quad + 2(\nabla_k u(t)(\nabla_l u(t) \cdot \nabla_k \nabla_l u(t)), |\Delta u(t)|^{q-2}\Delta u(t)) \\ &\quad - 2(u(t)(\nabla_l u(t) \cdot \nabla_k \nabla_l u(t)), |\Delta u(t)|^{q-2}\nabla_k \Delta u(t)) \\ &\quad - 2(u(t)(\nabla_l u(t) \cdot \nabla_k \nabla_l u(t)), \Delta u(t)\nabla_k (|\Delta u(t)|^{q-2})) \\ &\equiv I_1 + I_2 + I_3 + I_4. \end{aligned} \tag{3.2}$$

The first and second terms I_1, I_2 in (3.2) are dominated by the elliptic estimate in L^q (cf. [18]),

$$I_1 + I_2 \le \|\nabla u\|_\infty^2 \|\Delta u\|_q^q. \tag{3.3}$$

For the third term I_3, we again use the elliptic estimate to get

$$\begin{aligned} I_3 &\le \|u\|_\infty \int_{\mathbb{R}^n} |\nabla u||\nabla_k \nabla_l u| \cdot |\Delta u|^{q-2}|\nabla_k \Delta u|dx \\ &\le \|u\|_\infty \left(\int_{\mathbb{R}^n} |\nabla u|^2 |\Delta u|^{q-2} |\nabla_k \nabla_l u|^2 dx \right)^{1/2} \left(\int_{\mathbb{R}^n} |\Delta u|^{q-2} |\nabla_k \Delta u|^2 dx \right)^{1/2} \\ &\le C \int_{\mathbb{R}^n} |\nabla u|^2 |\Delta u|^{q-2} |\nabla_k \nabla_l u|^2 dx + \varepsilon \int_{\mathbb{R}^n} |\Delta u|^{q-2} |\nabla_k \Delta u|^2 dx \\ &\le C\|\nabla u\|_\infty^2 \|\Delta u\|_q^q + \frac{1}{2} \int_{\mathbb{R}^n} |\Delta u|^{q-2} |\nabla_k \Delta u|^2 dx. \end{aligned} \tag{3.4}$$

The last term I_4 can be dealt with in a similar manner:

$$
\begin{aligned}
I_4 &= \int_{\mathbb{R}^n} u_i \nabla_l u_j \nabla_k \nabla_l u_j \Delta u_i \nabla_k (|\Delta u|^2)^{(q-2)/2} dx \\
&= \frac{q-2}{2} \int_{\mathbb{R}^n} u_i \nabla_l u_j \nabla_k \nabla_l u_j \Delta u_i |\Delta u|^{q-4} \nabla_k (|\Delta u|^2) dx \\
&= (q-2) \int_{\mathbb{R}^n} u_i \nabla_l u_j \nabla_k \nabla_l u_j \Delta u_i |\Delta u|^{q-4} (\Delta u \cdot \nabla_k \Delta u) dx \\
&\le (q-2) \|u\|_\infty \left(\int_{\mathbb{R}^n} |\nabla u|^2 |\Delta u|^{q-2} |\nabla_k \nabla_l u|^2 dx \right)^{1/2} \left(\int_{\mathbb{R}^n} |\Delta u|^{q-2} |\nabla_k \Delta u|^2 dx \right)^{1/2} \\
&\le C \|\nabla u\|_\infty^2 \|\Delta u\|_q^q + \frac{1}{2} \int_{\mathbb{R}^n} |\Delta u|^{q-2} |\nabla_k \Delta u|^2 dx.
\end{aligned} \tag{3.5}
$$

On the other hand, the second term in the left-hand side of (3.2) is

$$
\begin{aligned}
&\int_{\mathbb{R}^n} \nabla_k \Delta u \cdot \nabla_k (|\Delta u|^{q-2} \Delta u) dx \\
&\quad = \int_{\mathbb{R}^n} |\Delta u|^{q-2} |\nabla_k \Delta u|^2 dx + \frac{1}{2} \int_{\mathbb{R}^n} \nabla_k |\Delta u|^2 \cdot \nabla_k |\Delta u|^{q-2} dx \\
&\quad = \int_{\mathbb{R}^n} |\Delta u|^{q-2} |\nabla_k \Delta u|^2 dx + \frac{q-2}{4} \int_{\mathbb{R}^n} |(|\Delta u|^2)^{(q-4)/4} \nabla_k |\Delta u|^2|^2 dx \\
&\quad = \int_{\mathbb{R}^n} |\Delta u|^{q-2} |\nabla_k \Delta u|^2 dx + \frac{4(q-2)}{q^2} \int_{\mathbb{R}^n} |\nabla (|\Delta u|^2)^{q/4}|^2 dx.
\end{aligned} \tag{3.6}
$$

Hence, by gathering estimates (3.3)–(3.5) and (3.6) and plugging them into (3.2), it follows that

$$
\begin{aligned}
&\frac{1}{q} \frac{d}{dt} \|\Delta u(t)\|_q^q + \frac{4(q-2)}{q^2} \|\nabla |\Delta u|^{q/2}\|_2^2 \\
&\qquad \le C \|\nabla u\|_\infty^2 \|\Delta u\|_q^q.
\end{aligned} \tag{3.7}
$$

Integration (3) over $[0, T]$ and the Young inequality imply

$$
\|\Delta u(t)\|_q^q \le \|\Delta u(0)\|_q^q + C(\varepsilon) \int_0^T \|\nabla u\|_\infty^2 \|\Delta u(\tau)\|_q^q d\tau. \tag{3.8}
$$

Noting the energy inequality (1.2), the logarithmic inequality (2.6) in Corollary 2.4 yields that, for $\gamma > n/q$ and $q > n$,

$$
\begin{aligned}
\|\nabla u\|_\infty &\le C(1 + \|\nabla u\|_{BMO} (\log^+ (\|\nabla u_+\|_{W^{1,q}} + \|u\|_\infty))^{1/2}) \\
&\le C(1 + \|\nabla u\|_{BMO} (\log^+ (\|\nabla u\|_{W^{1,q}} + 1))^{1/2}).
\end{aligned} \tag{3.9}
$$

Hence it follows from (3.8) and (3.9) that

$$
\begin{aligned}
\|\Delta u(t)\|_q^q &\le \|\Delta u_0\|_q^q + C \int_0^T \|\nabla u(\tau)\|_\infty^2 \|\Delta u(\tau)\|_q^q d\tau \\
&\le \|\Delta u_0\|_q^q + C \int_0^T \|\nabla u\|_{BMO}^2 (1 + \log^+ (\|\nabla u(\tau)\|_{W^{1,q}} + 1)^{\frac{1}{2}})^2 \|\Delta u(\tau)\|_q^q d\tau.
\end{aligned} \tag{3.10}
$$

Combining the energy inequality

$$\|\nabla u(t)\|_2^2 + 2\int_0^t \|\partial_t u(\tau)\|_2^2 d\tau \le \|\nabla u_0\|_2^2$$

with $\|u\|_\infty = 1$, we conclude by the Gronwall argument that

$$\|\nabla u(t)\|_{W^{1,q}}^q \le C\|\nabla u_0\|_{W^{1,q}}^q \exp\left\{C\exp\left(C\int_0^T (1+\|\nabla u\|_{BMO}^2)d\tau\right)\right\}.$$

This estimate ensures that the solution has regularity in $C((0,T];\dot{W}^{2,q})$ under the assumption (1.9). Since we have chosen that $q > n$, the Sobolev embedding implies that $\nabla u(t)$ is a continuous function in (x,t). A general argument for the harmonic heat flow gives the higher regularity. This completes the proof of Theorem 1.2. □

Acknowledgments. The author would like to thank Professor Masashi Misawa, Professor George Weiss, and Dr. Tôru Nakajima for stimulating discussion.

REFERENCES

[1] J. T. Beale, T. Kato, and A. Majda, *Remarks on the breakdown of smooth solutions for the 3-D Euler equations*, Comm. Math. Phys., 94 (1984), pp. 61–66.
[2] H. Beirão da Veiga, *A new regularity class for the Navier-Stokes equations in* $\mathbb{R}^n$, Chinese Ann. Math. Ser. B, 16, (1995), pp. 407–412.
[3] J. Bergh and J. Löfström, *Interpolation Spaces: An Introduction*, Grundlehren Math. Wiss. 223, Springer-Verlag, Berlin, New York, Heidelberg, 1976.
[4] H. Brezis and T. Gallouet, *Nonlinear Schrödinger evolution equations*, Nonlinear Anal., 4 (1980), pp. 677–681.
[5] H. Brezis and S. Wainger, *A note on limiting cases of Sobolev embedding and convolution inequalities*, Comm. Partial Differential Equations, 5 (1980), pp. 773–789.
[6] B. H. Qui, *Some aspects of weighted and non-weighted hardy spaces*, in The Study of Hardy Spaces and Several Variable Fourier Analysis by Real Analytical Methods, Sûrikaisekikenkyûsho Kôkyûroku 383, Kyoto University, Research Institute for Mathematical Sciences, Kyoto, Japan, 1980, pp. 38–56 (in Japanese).
[7] J.-Y. Chemin, *Perfect Incompressible Fluids*, Oxford Lecture Ser. Math. Appl. 14, Oxford University Press, New York, 1998.
[8] Y.-M. Chen and W.-Y. Ding, *Blow-up and global existence for heat flows of harmonic maps*, Invent. Math., 99 (1990), pp. 567–578.
[9] K.-C. Chang, W.-Y. Ding, and R. Ye, *Finite-time blow-up of the heat flow of harmonic maps from surfaces*, J. Differential Geom., 36 (1992), pp. 507–515.
[10] Y.-M. Chen and M. Struwe, *Existence and partial regularity results for the heat flow for harmonic maps*, Math. Z., 201 (1989), pp. 83–103.
[11] R. Coifman, P. L. Lions, Y. Meyer, and S. Semmes, *Compensated compactness and Hardy spaces*, J. Math. Pures Appl., 72 (1993), pp. 247–286.
[12] J.-M. Coron and J.-M. Ghidaglia, *Explosion en temps fini pour le flot des applications harmoniques*, C.R. Acad. Sci. Paris Ser. I Math., 308 (1989), pp. 339–344.
[13] J. Eells and J. H. Sampson, *Harmonic mappings of Riemannian manifolds*, Amer. J. Math., 86 (1964), pp. 109–160.
[14] H. Engler, *An alternative proof of the Brezis-Wainger inequality*, Comm. Partial Differential Equations, 14 (1989), pp. 541–544.
[15] L. C. Evans, *Partial regularity for stationary harmonic maps into spheres*, Arch. Ration. Mech. Anal., 116 (1991), pp. 101–113.
[16] M. Feldman, *Partial regularity for harmonic maps of evolution into spheres*, Comm. Partial Differential Equations, 19 (1994), pp. 761–790.
[17] Y. Giga, *Solutions for semilinear parabolic equations in* L^p *and regularity of weak solutions of the Navier-Stokes system*, J. Differential Equations, 62 (1986), pp. 186–212.
[18] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[19] F. Hélein, *Régularité des applications faiblement harmoniques entre une surface et une sphére*, C.R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 519–524.
[20] T. Kato and G. Ponce, *Commutator estimates and the Euler and Navier-Stokes equations*, Comm. Pure Appl. Math., 41 (1988), pp. 891–907.
[21] H. Kozono, T. Ogawa, and Y. Taniuchi, *The critical Sobolev inequalities in Besov spaces and regularity criterion to some semi-linear evolution equations*, Math. Z., 242 (2002), pp. 251–278.
[22] H. Kozono and Y. Taniuchi, *Limiting case of the Sobolev inequality in BMO with application to the Euler equations*, Comm. Math. Phys., 214 (2000), pp. 191–200.
[23] J. Leray, *Sur le mouvement d'un liquide visqeux emplissant l'espace*, Acta Math., 63 (1934), pp. 193–248.
[24] T. Ohyama, *Interior regularity of weak solutions of the time-dependent Navier-Stokes equation*, Proc. Japan Acad., 36 (1960), pp. 273–277.
[25] T. Ogawa and Y. Taniuchi, *Critial Sobolev Inequality and Uniqueness Problem to the Navier-Stokes Equations*, preprint, Kyushu University, Fukuoka, Japan.
[26] T. Ogawa and Y. Taniuchi, *On blow-up criteria of smooth solutions to the* 3*-D Euler equations in a bounded domain,* J. Differential Equations, in press (2003).
[27] G. Ponce, *Remarks on a paper by J.T. Beale, T. Kato and A. Majda*, Comm. Math. Phys., 98 (1985), pp. 349–353.
[28] J. Serrin, *On the interior regularity of weak solutions of the Navier-Stokes equations*, Arch. Ration. Mech. Anal., 9 (1962), pp. 187–195.
[29] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princton, NJ, 1970.
[30] R. S. Strichartz, *Bounded mean oscillation and Sobolev spaces*, Indiana Univ. Math. J., 29 (1980), pp. 539–558.
[31] M. Struwe, *Geometric evolution problems*, in Nonlinear Partial Differential Equations in Differential Geometry, R. Haardt and M. Wolf, eds., IAS/Park City Math. Ser. 2, AMS, Providence, RI, 1996, pp. 257–333.
[32] H. Triebel, *Theory of Function Spaces*, Monogr. Math. 78, Birkhäuser-Verlag, Basel, 1983.
[33] M. Vishik, *Hydrodynamics in Besov spaces*, Arch. Ration. Mech. Anal., 145 (1998), pp. 197–214.
[34] V. I. Yudovich, *Uniqueness theorem for the basic nonstationary problem in the dynamics of an ideal incompressible fluid*, Math. Res. Lett., 2 (1995), pp. 27–38.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1331–1355

A GAS-SOLID FREE BOUNDARY PROBLEM FOR A COMPRESSIBLE VISCOUS GAS*

FEIMIN HUANG†‡, AKITAKA MATSUMURA†, AND XIAODING SHI†§

Abstract. In this paper we propose a gas-solid free boundary problem for a one-dimensional model system of a compressible viscous gas associated with the inflow problem, where the solid constantly changes into the gas and produces the inflow of the gas on the free boundary. We first show the existence of the traveling wave solution and its asymptotic stability. We further discuss the case in which the asymptotic state of the solution is given by a combination of the traveling wave solution and rarefaction wave.

Key words. gas-solid free boundary, compressible viscous gas, traveling wave, rarefaction wave

AMS subject classification. 35L65

PII. S0036141002403730

1. Introduction. We consider the one-dimensional barotropic motion of compressible viscous gas and study the situation where the gas is adjacent to the solid by the free boundary $\tilde{x} = X(t)$. In what follows we use the *Eulerian* coordinates $(\tilde{x}, t)$ to represent the spatial and time variables. We assume that the part $\tilde{x} < X(t)$ is filled by the solid whose density $\bar{\rho}$ is a given positive constant and whose velocity $\bar{u}$ is zero, and that the part $\tilde{x} > X(t)$ is filled by the gas with the density $\tilde{\rho}(\tilde{x}, t)$ and velocity $\tilde{u}(\tilde{x}, t)$ satisfying the conservation of mass

$$\tilde{\rho}_t + (\tilde{\rho}\tilde{u})_{\tilde{x}} = 0 \tag{1.1}$$

and the conservation of momentum

$$(\tilde{\rho}\tilde{u})_t + (\tilde{\rho}\tilde{u}^2 + \tilde{p} - \mu\tilde{u}_{\tilde{x}})_{\tilde{x}} = 0, \tag{1.2}$$

where the pressure $\tilde{p}$ is a given function of $\tilde{\rho}$, and the viscosity coefficient μ is a given positive constant. Here we should note that in the solid part, $\tilde{\rho} = \bar{\rho}$ and $\tilde{u} = \bar{u} = 0$ also satisfy (1.1) and (1.2). On the free boundary, so that the conservation of mass holds, we assume the Rankine–Hugoniot condition

$$-\frac{dX(t)}{dt}(\tilde{\rho} - \bar{\rho}) + (\tilde{\rho}\tilde{u} - \bar{\rho}\bar{u}) = 0 \quad \text{on } \tilde{x} = X(t),$$

which implies

$$\frac{dX(t)}{dt} = \frac{\tilde{\rho}\tilde{u}}{\tilde{\rho} - \bar{\rho}}. \tag{1.3}$$

*Received by the editors March 8, 2002; accepted for publication (in revised form) October 24, 2002; published electronically May 12, 2003.
http://www.siam.org/journals/sima/34-6/40373.html

†Department of Mathematics, Graduate School of Science, Osaka University, Osaka 560-0045, Japan (fhuang@mail.amt.ac.cn, akitaka@ist.osaka-u.ac.jp, shixd@mail.buct.edu.cn). The work of the first author was supported in part by the JSPS Research Fellowship for foreign researchers and Grant-in-Aid P-00269 for JSPS from the Ministry of Education, Science, Sports and Culture of Japan.

‡Institute of Applied Mathematics, AMSS, Academia Sinica, Beijing 100080, China.
§Department of Mathematics, Graduate School of Science, Beijing University of Technology and Chemical, Beijing 100029, China.

Depending on the sign of $X'(t)$ and also on whether $\tilde{\rho}$ is more or less than $\bar{\rho}$, we can consider various situations. Among them we study here the situation

$$\tilde{\rho} < \bar{\rho}, \qquad \frac{dX(t)}{dt} < 0. \tag{1.4}$$

It is noted that the first condition is physically natural because the density of the gas changed in phase from the solid is usually much less than that of the solid, and the second condition makes us focus on the phase transition process from the solid to the gas. In this situation, since the density in front of the free boundary is more than that in back of the boundary, the discontinuity of the solution at the boundary is not a classical shock but a *detonation*-type discontinuity. By (1.3) we also know the velocity $\tilde{u}$ on the boundary is positive, which means the gas constantly flows in from the free boundary, so this situation is closely related to the inflow problem on the half space for (1.1), (1.2) (cf. [5, 19]). We discuss this matter more later. Since we need another boundary condition on the density in this inflow situation, we assume

$$\tilde{\rho} = \rho_b \quad \text{on } \tilde{x} = X(t), \qquad \tilde{p}(\rho_b) = \bar{p} := \tilde{p}(\bar{\rho}), \tag{1.5}$$

where $\rho_b\,(< \bar{\rho})$ is a given positive constant, and the second assumption means the pressure is continuous across the interface. We also assume

$$\tilde{p}(\tilde{\rho}) > 0, \quad \tilde{p}'(\tilde{\rho}) > 0, \quad \tilde{p}''(\tilde{\rho}) \geq 0 \qquad \text{in a neighborhood of } \tilde{\rho} = \rho_b, \tag{1.6}$$

which means that once the solid changes to the gas it follows the standard property of pressure-density relation as barotropic gas. Since we need another boundary condition to determine the movement of the free boundary, the data of the density, and momentum on the boundary, we also assume another Rankine–Hugoniot condition for the conservation of momentum

$$-\frac{dX}{dt}\tilde{\rho}\tilde{u} + \tilde{\rho}\tilde{u}^2 + \tilde{p} - \mu\tilde{u}_{\tilde{x}} - \bar{p} = 0 \quad \text{on } \tilde{x} = X(t). \tag{1.7}$$

By (1.3), (1.5), and (1.7), we have a nonlinear Neumann condition for the velocity

$$\mu\tilde{u}_{\tilde{x}} = \frac{\bar{\rho}\rho_b}{\bar{\rho} - \rho_b}\tilde{u}^2. \tag{1.8}$$

We know that to be more physical, especially to unify the arguments of fluid dynamical aspects and that of Stephan problems, we should further take into account the conservation of energy and its Rankine–Hugoniot condition with a jump of energy structure. However, even for 2×2 systems like (1.1), (1.2) there have been no mathematical results under the free boundary condition like (1.3) because of various difficulties. In all previous works (e.g., Kazhikhov [10], Nagasawa [20]) they assume

$$\frac{dX(t)}{dt} = \tilde{u}, \tag{1.9}$$

which means the particles on the free boundary always stay on the boundary. In this case, if we introduce the Lagrangian mass coordinates, we can reformulate the problem to that with the fixed boundary. On the other hand, in our case (1.3) we cannot do it, which gives us serious difficulty. We believe that to investigate a simple mathematical model as (1.1), (1.2) under the conditions (1.3), (1.8) is a quite meaningful one-step to overcoming mathematical difficulties and proceeding to more general problems. We

should note that Kaliev and Kazhikhov [7] showed the existence and uniqueness of the local solution for a free boundary value problem which does not satisfy (1.9). But in their case they assumed that there is no discontinuity of the density. That is, $\tilde{\rho} = \rho_b$ on the boundary. Thus we propose and concentrate on the following system in the *Eulerian* coordinates:

$$
\begin{cases}
\tilde{\rho}_t + (\tilde{\rho}\tilde{u})_{\tilde{x}} = 0 & \text{in } \tilde{x} > X(t), \\
(\tilde{\rho}\tilde{u})_t + (\tilde{\rho}\tilde{u}^2 + \tilde{p})_{\tilde{x}} = \mu \tilde{u}_{\tilde{x}\tilde{x}} & \text{in } \tilde{x} > X(t), \\
\tilde{\rho}(X(t), t) = \rho_b, & \\
\mu \tilde{u}_{\tilde{x}}(X(t), t) = \dfrac{\rho_b \bar{\rho}}{\bar{\rho} - \rho_b} \tilde{u}^2(X(t), t), & \\
\dfrac{dX(t)}{dt} = \dfrac{\rho_b}{\rho_b - \bar{\rho}} \tilde{u}(X(t), t) & X(0) = 0, \\
(\tilde{\rho}, \tilde{u})|_{(+\infty, t)} = (\rho_+, u_+), & \\
(\tilde{\rho}, \tilde{u})|_{t=0} = (\rho_0, u_0). &
\end{cases}
\tag{1.10}
$$

It is worthwhile to point out that in (1.10) the density $\tilde{\rho}$ must be imposed on the boundary because the boundary is moving to the left, while the fluid particles on the boundary are moving to the right, so that characteristics for the $\tilde{\rho}$ equation come *out* of the boundary in forward time.

In this paper we first investigate the existence of the traveling wave solution of (1.10) and establish its asymptotic stability. We further discuss the case in which the asymptotic state of the solution is expected to consist of the traveling wave solution and rarefaction wave.

We now state the relations of (1.10) to the inflow problem with the fixed boundary and recall the results on it. We consider the case $X(t) = \bar{s}t (\bar{s} < 0)$. Let $y = \tilde{x} - \bar{s}t$, $z(y, t) = \tilde{u}(\tilde{x}, t) - \bar{s}$, $\rho(y, t) = \tilde{\rho}(\tilde{x}, t)$; then (1.3) and (1.5) yield the following boundary condition:

$$
\rho|_{y=0} = \rho_b > 0, \quad z|_{y=0} = -\frac{\bar{\rho}\bar{s}}{\rho_b} =: u_b > 0.
$$

Thus the system (1.10) is changed into

$$
\begin{cases}
\rho_t + (\rho z)_y = 0 & \text{in } y > 0, \\
(\rho z)_t + (\rho z^2 + p)_y = \mu z_{yy} & \text{in } y > 0, \\
\rho(0, t) = \rho_b > 0, & \\
z(0, t) = u_b > 0, & \\
(\rho, z)|_{(+\infty, t)} = (\rho_+, u_+ - \bar{s}), & \\
(\rho, z)|_{t=0} = (\rho_0, u_0 - \bar{s}), &
\end{cases}
\tag{1.11}
$$

where $p = \tilde{p}(\rho(y, t))$.

The problem (1.11) is called the inflow problem with the fixed boundary. In this situation, a new wave, denoted by the boundary layer solution, or BL-solution, appears in the solutions due to the presence of a boundary. Matsumura [13] classified all possible large time behaviors of the solutions in terms of the boundary values. Matsumura and Nishihara [19] and we [5] have proved that if the boundary values are in the subsonic region, the solution tends to the superposition of a BL-solution and

a rarefaction wave, and the superposition of a BL-solution and a viscous shock wave under some smallness conditions. Shi [23] studied the rarefaction wave case when the boundary values are in the supersonic region. We refer the readers to [5, 13, 19, 23].

We now concentrate on the free boundary problem (1.10). We here transform (1.10) to the problem in the Lagrangian coordinate:

$$
\begin{cases}
v_t - u_x = 0, & x > x(t), \quad t > 0,\\
u_t + p(v)_x = \mu\left(\dfrac{u_x}{v}\right)_x, & x > x(t), \quad t > 0,\\
v(x(t),t) = v_b,\\
\mu u_x(x(t),t) = \dfrac{v_b}{v_b - \bar{v}} u^2(x(t),t),\\
\dfrac{dx(t)}{dt} = \dfrac{1}{\bar{v} - v_b} u(x(t),t), \quad x(0) = 0,\\
(v,u)|_{(+\infty,t)} = (v_+, u_+) = \left(\dfrac{1}{\rho_+}, u_+\right),\\
(v,u)|_{t=0} = (v_0, u_0), \quad v_0(0) = v_b,
\end{cases}
\tag{1.12}
$$

where $v = 1/\rho$. The transformation $(\tilde{x}, t) \to (x, t)$ is given by

$$
\begin{cases}
\dfrac{\partial \tilde{x}(x,t)}{\partial t} = \tilde{u}(\tilde{x}(x,t),t), \quad t > 0, \quad x > 0\\
\tilde{x}(x,0) = \tilde{x}_0(x),
\end{cases}
$$

with

$$
\int_0^{\tilde{x}_0(x)} \tilde{\rho}(r,0)dr = x,
$$

where $(\tilde{x}, t) \in \sum_1 = \{(\tilde{x}, t); \tilde{x} > \tilde{x}(0, t)\}$, and by

$$
\begin{cases}
\dfrac{\partial \tilde{x}(x,t)}{\partial t} = \tilde{u}(\tilde{x}(x,t),t), \quad t > t_0(x), \quad x > 0\\
\tilde{x}(x, t_0(x)) = X(t_0(x)),
\end{cases}
$$

with

$$
x = \int_0^{X(t_0(x))} \rho_b dr - \int_0^{t_0(x)} \rho_b \tilde{u}(X(t),t)dt = \bar{\rho} X(t_0(x)),
$$

where $(\tilde{x}, t) \in \sum_2 = \{(\tilde{x}, t); X(t) < \tilde{x} < \tilde{x}(0, t)\}$. From the definition, we have

$$
x \geq x(t) = \bar{\rho} X(t)
$$

and

$$
\int_{\tilde{x}(0,t)}^{\tilde{x}(x,t)} \tilde{\rho}(r,t)dr = x
$$

for $(\tilde{x}, t) \in \sum_i (i = 1, 2)$. Hence, for $f(x, t) = \tilde{f}(\tilde{x}(x, t), t)$,

$$
\frac{\partial}{\partial t} f(x,t) = \left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial \tilde{x}}\right)\tilde{f}(\tilde{x}(x,t),t), \quad \frac{\partial}{\partial x} f(x,t) = v\frac{\partial}{\partial \tilde{x}}\tilde{f}(\tilde{x}(x,t),t),
$$

which gives (1.12).

We now seek a traveling wave solution of (1.12). The argument of section 2 shows that for any given $0 < \bar{v} < v_b < v_+$, if

$$u_+ = (v_+ - \bar{v})^{\frac{1}{2}}(p(v_b) - p(v_+))^{\frac{1}{2}}, \quad u_b = \frac{\bar{v} - v_b}{\bar{v} - v_+}u_+, \quad \bar{s} = \frac{u_b}{\bar{v} - v_b}, \tag{1.13}$$

there exists a unique traveling wave solution $(V, U)(\xi)$, $\xi = x - \bar{s}t$, satisfying

$$\begin{cases} -\bar{s}V' - U' = 0, \\ -\bar{s}U' + p(V)' = \mu\left(\dfrac{U'}{V}\right)', \end{cases} \tag{1.14}$$

with

$$\begin{cases} V(0) = v_b, \quad U(0) = u_b, \quad \mu U'(0) = v_b(v_b - \bar{v})\bar{s}^2, \\ V(+\infty) = v_+, \quad U(+\infty) = u_+, \end{cases} \tag{1.15}$$

provided that $u_+^2 < (v_+ - \bar{v})^2|p'(v_+)|$. We call (V, U) the traveling wave solution. Therefore for any $0 < \bar{v} < v_b < +\infty$, the traveling wave solution curve with the parameter $\bar{v}$ through the point $(v_b, 0)$ in the phase plane is defined by

$$\begin{aligned} &TW_{\bar{v}}(v_b, 0) \\ &= \left\{(v, u); u = (v - \bar{v})^{\frac{1}{2}}(p(v_b) - p(v))^{\frac{1}{2}}, u^2 < (v - \bar{v})^2|p'(v)|, v > v_b\right\}. \end{aligned} \tag{1.16}$$

When $(v_+, u_+) \in TW_{\bar{v}}(v_b, 0)$, the solution of (1.12) is expected to tend to the traveling wave solution satisfying (1.13)–(1.15) as t tends to infinity.

On the other hand, it is known that the 2-rarefaction curve $R_2(v_b, 0)$ through the point $(v_b, 0)$ is

$$R_2(v_b, 0) = \left\{(v, u); u = -\int_{v_b}^{v} \lambda_2(s)ds, v < v_b\right\} \tag{1.17}$$

and $R_2(v_p, u_p)$ through the point (v_p, u_p) is

$$R_2(v_p, u_p) = \left\{(v, u); u = u_p - \int_{v_p}^{v} \lambda_2(s)ds, v < v_p\right\}, \tag{1.18}$$

where $\lambda_2 = \sqrt{-p'(v)}$ is the second characteristic speed of the corresponding hyperbolic system without viscosity and

$$u_p = (v_p - \bar{v})^{\frac{1}{2}}(p(v_b) - p(v_p))^{\frac{1}{2}}, \quad u_p^2 = (v_p - \bar{v})^2|p'(v_p)|.$$

Then let us define $R_2TW_{\bar{v}}(v_b, 0)$ as the domain surrounded by curves $TW_{\bar{v}}(v_b, 0)$, $R_2(v_b, 0)$, $R_2(v_p, u_p)$, and the u-axis. When (v_+, u_+) is in $R_2TW_{\bar{v}}(v_b, 0)$, the solution of (1.12) is expected to tend to the superposition of a traveling wave solution and a 2-rarefaction wave as t tends to infinity.

Our aim in this present paper is to investigate the stability of the traveling wave solution and of a superposition of the traveling wave solution and the 2-rarefaction wave. Our results are, roughly speaking, as follows.

(I) If $(v_+, u_+) \in TW_{\bar{v}}(v_b, 0)$, then the traveling wave solution is stable, provided that $|v_+ - v_b|$ is small. The precise statement is given in Theorem 4.1 below.

(II) If $(v_+, u_+) \in R_2TW_{\bar{v}}(v_b, 0)$, then there exists $(v_*, u_*) \in TW_{\bar{v}}(v_b, 0)$ such that $(v_+, u_+) \in R_2(v_*, u_*)$, and the superposition of the traveling wave solution connecting (v_b, u_b) with (v_*, u_*) and the 2-rarefaction wave connecting (v_*, u_*) with (v_+, u_+) is stable provided that $|v_* - v_b|$ is small. That is, the traveling wave solution is necessary to be weak; however, the 2-rarefaction wave is not necessarily weak. The precise statement is given in Theorem 5.3 below.

Remark. Although we only study the case (1.4) in this paper, the case $\tilde{\rho} > \bar{\rho}$, $\frac{dX(t)}{dt} < 0$ could be treated by the same method. We note that in this case both the boundary and the fluid are moving to the left. However, the fluid speed is faster than that of the boundary. Therefore, similar to the case (1.4), the characteristics for the $\tilde{\rho}$ equation still come *out* of the boundary in forward time and the density $\tilde{\rho}$ must be imposed on the boundary.

It is interesting to compare our results with those of the traveling wave with shock profile. For the stability of the viscous shock wave, it is known that the scalar equation has been extensively investigated (cf. [6, 21, 22]). Studies on systems began with the independent works of Goodman [3] and Matsumura and Nishihara [15] with zero mass. The generic initial perturbations were investigated by Liu [11] and Szepessy and Xin [24]. Unlike the viscous shock wave, since the traveling wave is unique here, it is not necessary to locate which of its translates the perturbed solution converges to. Therefore we do not need any hypothesis that states that the perturbations in Theorems 4.1 and 5.3 have zero mass.

Our method is based on the energy estimates and a new approximation to the rarefaction wave. It is noted that the new approximation could be applied to extend the results of [19]—where a stability theorem for the superposition of a BL-solution and a weak rarefaction wave with the fixed boundary was obtained by Matsumura and Nishihara—to the strong rarefaction wave.

Our paper is organized as follows. In sections 2–4, we focus on case I. Precisely speaking, in section 2, we study the existence of the traveling wave solution. In section 3, we reformulate the original problem to a new initial boundary value problem and prove the local existence of the solution. In section 4, we establish the a priori estimates and then prove the stability of the traveling wave solution. In section 5, case II is treated.

Notation. Throughout this paper, several positive generic constants are denoted by c, C without confusion. For function spaces, $L^p(\Omega), 1 \le p \le \infty$, denotes a usual Lebesgue space on $\Omega \subset R = (-\infty, \infty)$ with its norm

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p dx \right)^{\frac{1}{p}}, \quad 1 \le p < \infty, \quad \|f\|_{L^\infty(\Omega)} = \sup_\Omega |f(x)|. \tag{1.19}$$

$H^l(\Omega)$ denotes the lth order Sobolev space with its norm

$$\|f\|_l = \left(\sum_{j=0}^{l} \|\partial_x^j f\|^2 \right)^{\frac{1}{2}}, \quad \text{when } \|\cdot\| := \|\cdot\|_{L^2(\Omega)}. \tag{1.20}$$

The domain Ω will often be abbreviated to avoid confusion.

2. Existence of the traveling wave solution. In this section we investigate the existence of the traveling wave solution $(V,U)(\xi)$, $\xi = x - \bar{s}t$, satisfying

$$\begin{cases} -\bar{s}V' - U' = 0, \qquad ' = d/d\xi, \quad \xi = x - \bar{s}t, \\ -\bar{s}U' + p(V)' = \mu\left(\dfrac{U'}{V}\right)', \\ V(0) = v_b, \quad U(0) = u_b, \\ \mu U'(0) = v_b(v_b - \bar{v})\bar{s}^2, \\ V(+\infty) = v_+, \quad U(+\infty) = u_+, \end{cases} \tag{2.1}$$

where

$$\bar{s} = \frac{u_b}{\bar{v} - v_b} < 0. \tag{2.2}$$

We have the following existence result.

LEMMA 2.1. *For any given $\bar{v}, v_b$, and v_+ with $0 < \bar{v} < v_b < v_+$, when*

$$u_b = \frac{\bar{v} - v_b}{\bar{v} - v_+}u_+, \quad u_+ = (v_+ - \bar{v})^{\frac{1}{2}}(p(v_b) - p(v_+))^{\frac{1}{2}}, \tag{2.3}$$

there exists a unique solution $(V,U)(\xi)$ to (2.1) and (2.2) satisfying $0 < v_b < V(\xi) < v_+, V' > 0$, provided that $u_+^2 < (v_+ - \bar{v})^2|p'(v_+)|$. Furthermore, fix $\bar{v}$ and v_b, and let $v_+ - v_b = \delta$; then there exists a positive constant $\delta_0 > 0$ such that, for any $\delta \le \delta_0$,

$$|V(\xi) - v_+| = O(\delta)e^{-\frac{c}{\sqrt{\delta}}\xi} \tag{2.4}$$

and

$$u_+, \bar{s}, u_b = O(\delta^{\frac{1}{2}}), \quad V' = O(\delta^{\frac{1}{2}})e^{-\frac{c}{\sqrt{\delta}}\xi}. \tag{2.5}$$

Proof. When (V,U) exists, the integration of (2.1) with respect to ξ yields

$$\begin{cases} \bar{s}V + U = \bar{s}v_b + u_b = \bar{s}v_+ + u_+, \\ \mu\dfrac{U'}{V} = p(V) + \bar{s}^2(V - v_+) - p(v_+). \end{cases} \tag{2.6}$$

Let $\xi = 0$; then (2.3) holds. By (2.6) we have the ordinary equation

$$\begin{cases} -\bar{s}\mu\dfrac{V'}{V} = \mathcal{F}(V) := p(V) + \bar{s}^2(V - v_+) - p(v_+), \\ V(0) = v_b, \quad V(+\infty) = v_+. \end{cases} \tag{2.7}$$

To the contrary, since

$$u_+^2 < (v_+ - \bar{v})^2|p'(v_+)| \tag{2.8}$$

implies $\mathcal{F}(V) > 0$ for $v_b < V < v_+$, it is easy to see there exists a unique solution (V,U) of (2.1). Furthermore, (2.3) and (2.7) imply that the solution satisfies (2.4) and (2.5) if $v_+ - v_b$ is suitably small. Thus Lemma 2.1 is proved.

3. Local existence of the solution. In this section, we first reformulate (1.12) to a new initial-boundary value problem and then prove the local existence of the solution.

Assume that $(v_+, u_+) \in TW_{\bar{v}}(v_b, 0)$. Then Lemma 2.1 gives a unique solution $(V, U)(\xi), \xi = x - \bar{s}t \geq 0$, satisfying

$$\begin{cases} -\bar{s}V' - U' = 0, \\ -\bar{s}U' + p(V)' = \mu\left(\dfrac{U'}{V}\right)', \\ V(0) = v_b, U(0) = u_b, \\ \mu U'(0) = v_b(v_b - \bar{v})\bar{s}^2, \\ (V, U)|_{(+\infty)} = (v_+, u_+), \end{cases} \tag{3.1}$$

where $\bar{s} = \frac{u_b}{\bar{v} - v_b} < 0$ and u_b, u_+ satisfies (2.3). To investigate the free boundary problem (1.12), we consider the coordinate transformation

$$t = t, \quad y = x - x(t), \tag{3.2}$$

where $x(t) = \bar{s}t + \gamma(t)$. By (3.2), we rewrite (1.12) as

$$\begin{cases} v_t - (\bar{s} + \gamma'(t))v_y - u_y = 0, & y > 0, \quad t > 0, \\ u_t - (\bar{s} + \gamma'(t))u_y + p(v)_y = \mu\left(\dfrac{u_y}{v}\right)_y, & y > 0, \quad t > 0, \\ v(0, t) = v_b, \\ \mu u_y(0, t) = \dfrac{v_b}{v_b - \bar{v}}u^2(0, t), \\ \dfrac{d\gamma(t)}{dt} = \dfrac{1}{\bar{v} - v_b}u(0, t) - \bar{s}, \quad \gamma(0) = 0, \\ (v, u)|_{(+\infty, t)} = (v_+, u_+), \\ (v, u)(y, 0) = (v_0, u_0)(y), \quad y > 0. \end{cases} \tag{3.3}$$

On the other hand, we take the traveling wave solution as the form $(V, U)(y)$. Thus, by (3.1) (V, U) satisfies

$$\begin{cases} -\bar{s}V_y - U_y = 0, & y > 0, \\ -\bar{s}U_y + p(V)_y = \mu\left(\dfrac{U_y}{V}\right)_y, & y > 0, \\ V(0) = v_b, U(0) = u_b, \\ \mu U_y(0) = v_b(v_b - \bar{v})\bar{s}^2, \\ (V, U)|_{(+\infty)} = (v_+, u_+). \end{cases} \tag{3.4}$$

We now put the perturbation $(\phi, \psi)(y, t)$ by

$$(v, u)(y, t) = (V, U)(y) + (\phi, \psi)(y, t), \tag{3.5}$$

so that the reformulated problem is

$$
\begin{cases}
\phi_t - (\bar{s}+\gamma'(t))\phi_y - \psi_y = \gamma'(t)V'(y), & y>0, \quad t>0,\\
\psi_t - (\bar{s}+\gamma'(t))\psi_y + (p(V+\phi)-p(V))_y & \\
-\mu\left(\dfrac{\psi_y}{V+\phi} + \dfrac{\bar{s}V'(y)\phi}{V(V+\phi)}\right)_y = -\bar{s}\gamma'(t)V'(y), & y>0, \quad t>0,\\
\phi(0,t)=0, & \\
\mu\psi_y(0,t) = \dfrac{2v_bu_b}{v_b-\bar{v}}\psi(0,t) + \dfrac{v_b}{v_b-\bar{v}}\psi^2(0,t), & \\
\gamma'(t) = \dfrac{1}{\bar{v}-v_b}\psi(0,t), \quad \gamma(0)=0, & \\
(\phi,\psi)|_{(+\infty,t)} = (0,0), & \\
(\phi,\psi)(y,0) = (\phi_0,\psi_0) = (v_0-V, u_0-U)(y). &
\end{cases} \tag{3.6}
$$

We shall combine the local existence and the a priori estimates to investigate the stability of the traveling wave solution. The a priori estimates will be treated in the next section. Here we only study the problem of local existence.

The solution space is

$$
\begin{aligned}
X_{m,M}(0,T) = \Big\{ & \phi \in C(0,T;H_0^1), \psi \in C(0,T;H^1) | \phi_y \in L^2(0,T;L^2),\\
& \psi_y \in L^2(0,T;H^1), \text{with } \sup_{[0,T]} \|(\phi,\psi)(t)\|_1 \le M,\\
& \inf_{R_+\times[0,T]} (V+\phi)(y,t) \ge m \Big\}
\end{aligned} \tag{3.7}
$$

for some positive constants m, M.

PROPOSITION 3.1 (local existence). *For any given $0<\bar{v}<v_b<+\infty$, there exist positive constants δ_0 and M_0 such that if $v_+ - v_b = \delta \le \delta_0$, $\phi_0 \in H_0^1$, $\psi_0 \in H^1$, $\|(\phi_0,\psi_0)\|_1 \le M(\le \frac{M_0}{b})$, and $\inf_{\mathbb{R}_+}(V+\phi_0) \ge m$, then there exists a positive constant $T_0 = T(m,M_0)$ such that there exists a unique solution $(\phi(y,t),\psi(y,t),\gamma(t))$ to (3.6) satisfying*

$$
\phi(y,t), \psi(y,t) \in X_{\frac{m}{2},bM}(0,T_0), \quad \gamma(t) \in C^1([0,T_0]),
$$

where $b = 3(1+2\sqrt{\frac{v_b}{\mu}} + \sqrt{\frac{2v_b}{v_b-\bar{v}}})$.

Proof. We first consider the characteristic equation of $(3.6)_1$. When $y \ge -x(t)$ ($x(t) = \bar{s}t + \gamma(t)$), the characteristic starts from the y-axis. That is, for any $\bar{x}_0 \ge 0$, the characteristic $y(\bar{x}_0,t)$ from $(\bar{x}_0,0)$ satisfies

$$
\begin{cases}
\dfrac{dy(\bar{x}_0,t)}{dt} = -\bar{s} - \gamma'(t),\\
y(\bar{x}_0,0) = \bar{x}_0,
\end{cases} \tag{3.8}
$$

which yields

$$
y(\bar{x}_0,t) = -x(t) + \bar{x}_0. \tag{3.9}
$$

Thus, ϕ has the explicit form

$$\phi(y,t)=\phi_0(\bar{x}_0)+\int_0^t \psi_y(\bar{x}_0-x(\tau),\tau)d\tau + \int_0^t \gamma'(\tau)V'(\bar{x}_0-x(\tau))d\tau, \quad \bar{x}_0=y+x(t). \tag{3.10}$$

In the same way, when $0\le y\le -x(t)$, the characteristic starts from the t-axis. It is noted that the inverse function of $x(t)$ exists if $|\gamma'(t)|=|\frac{1}{\bar{v}-v_b}\psi(0,t)|\le c\|\psi(t)\|_1$ is small. Thus, by $(3.6)_1$ and the boundary condition $\phi(0,t)=0$, we have

$$\phi(y,t)=\int_{\bar{t}_0}^t \psi_y(x(\bar{t}_0)-x(\tau),\tau)d\tau + \int_{\bar{t}_0}^t \gamma'(\tau)V'(x(\bar{t}_0)-x(\tau))d\tau, \quad \bar{t}_0=x^{-1}(y+x(t)). \tag{3.11}$$

On the other hand, $(3.6)_2$ is regarded as the initial-boundary value problem for the parabolic equation of ψ:

$$\begin{cases} \psi_t-\mu(\dfrac{\psi_y}{V+\phi})_y=g:=g(\phi,\phi_y,\psi(0,t),\psi_y) \\ \mu\psi_y(0,t)=\dfrac{2v_bu_b}{v_b-\bar{v}}\psi(0,t)+\dfrac{v_b}{v_b-\bar{v}}\psi^2(0,t), \\ \psi|_{t=0}=\psi_0, \end{cases} \tag{3.12}$$

where

$$\begin{aligned} & g(\phi,\phi_y,\psi(0,t),\psi_y) \\ = & -(p(V+\phi)-p(V))_y+\left(\bar{s}+\frac{1}{\bar{v}-v_b}\psi(0,t)\right)\psi_y \\ & +\mu\left(\frac{U_y}{V+\phi}-\frac{U_y}{V}\right)_y-\frac{\bar{s}}{\bar{v}-v_b}\psi(0,t)V'(y). \end{aligned}$$

By virtue of the boundary condition of (3.6), $\gamma(t)$ has the explicit form

$$\gamma(t)=\int_0^t \frac{1}{\bar{v}-v_b}\psi(0,\tau)d\tau. \tag{3.13}$$

We now approximate $\phi_0\in H^1_0, \psi_0\in H^1$ by $\phi_{0k}\in H^3\cap H^1_0, \psi_{0k}\in H^3$ such that

$$(\phi_{0k},\psi_{0k})\to(\phi_0,\psi_0) \quad \text{strongly} \quad \text{in} \quad H^1 \tag{3.14}$$

as $k\to\infty$ and $\|(\phi_{0k},\psi_{0k})\|_1\le\frac{3}{2}M$, $\inf_{R_+}(V+\phi_{0k})\ge\frac{2}{3}m$ hold for any k.

We will use the iteration method to prove Proposition 3.1. We define the sequence $\{(\phi_k^{(n)}(y,t),\psi_k^{(n)}(y,t),\gamma_k^{(n)}(t))\}$ for each k so that

$$(\phi_k^{(0)},\psi_k^{(0)})(y,t)=(\phi_{0k},\psi_{0k})(y), \quad \gamma_k^{(0)}(t)=\int_0^t \frac{1}{\bar{v}-v_b}\psi_{0k}(0)d\tau \tag{3.15}$$

and for a given $((\phi_k^{(n-1)}, \psi_k^{(n-1)})(y,t), \gamma_k^{(n-1)}(t))$, $\psi_k^{(n)}$ is a solution to

(3.16)
$$\begin{cases} \psi_{kt}^{(n)} - \mu\left(\dfrac{\psi_{ky}^{(n)}}{V+\phi_k^{(n-1)}}\right)_y = g^{(n-1)} = g\left(\phi_k^{(n-1)}, \phi_{ky}^{(n-1)}, \psi_k^{(n-1)}(0,t), \psi_{ky}^{(n-1)}\right), \\ \mu\psi_{ky}^{(n)}(0,t) = \dfrac{2v_b u_b}{v_b-\bar v}\psi_k^{(n)}(0,t) + \dfrac{v_b}{v_b-\bar v}[\psi_k^{(n)}(0,t)]^2, \\ \psi_k^{(n)}|_{t=0} = \psi_{0k}, \end{cases}$$

$$\gamma_k^{(n)}(t) = \int_0^t \frac{1}{\bar v - v_b}\psi_k^{(n)}(0,\tau)d\tau, \tag{3.17}$$

and

$$\phi_k^{(n)}(y,t) = \begin{cases} \displaystyle\int_{\bar t_k^{(n-1)}}^t \psi_{ky}^{(n)}(x_k^{(n-1)}(\bar t_k^{(n-1)}) - x_k^{(n-1)}(\tau), \tau)d\tau \\ \displaystyle + \int_{\bar t_k^{(n-1)}}^t [\gamma_k^{(n-1)}(\tau)]' V'(x_k^{(n-1)}(\bar t_k^{(n-1)}) - x_k^{(n-1)}(\tau))d\tau \\ \qquad \text{if} \quad 0 \le y \le -x_k^{(n-1)}(t), \\ \phi_{0k}(\bar x_k^{(n-1)}) + \displaystyle\int_0^t \psi_{ky}^{(n)}(\bar x_k^{(n-1)} - x_k^{(n-1)}(\tau), \tau)d\tau \\ \displaystyle + \int_0^t [\gamma_k^{(n-1)}(\tau)]' V'(\bar x_k^{(n-1)} - x_k^{(n-1)}(\tau))d\tau, \\ \qquad \text{if} \quad y \ge -x_k^{(n-1)}(t), \end{cases} \tag{3.18}$$

where

$$x_k^{(n-1)}(t) = \bar s t + \gamma_k^{(n-1)}(t),$$

$$\bar t_k^{(n-1)} = (x_k^{(n-1)})^{-1}(y + x_k^{(n-1)}(t)),$$

$$\bar x_k^{(n-1)} = y + x_k^{(n-1)}(t).$$

We now assume M_0 is small. By the principle of contraction mapping, it is easy to prove there exists a positive time $t_0(m, M_0) \ll 1$ such that if $g^{(n-1)} \in C(0, t_0; H^2)$ and $\psi_{0k} \in H^3$, there exists a unique local solution $\psi_k^{(n)}$ to (3.16) satisfying

$$\psi_k^{(n)} \in C(0,t_0;H^3) \cap C^1(0,t_0;H^1) \cap L^2(0,t_0;H^4)$$

and $|\psi_k^{(n)}(y,t)| \le CM_0$.

Thus, if $(\phi_k^{(n-1)}, \psi_k^{(n-1)}) \in X_{\frac{1}{2}m,bM}$, multiplying (3.16) by $\psi_k^{(n)}$ and integrating it over R_+, we have

$$\begin{aligned} &\|\psi_k^{(n)}(t)\|_t^2 + \frac{\mu}{v_+}\|\psi_{ky}^{(n)}\|^2 + \frac{4v_bu_b}{v_b-\bar v}[\psi_k^{(n)}(0,t)]^2 \\ \le\ & 2\int_0^\infty |g^{(n-1)}\psi_k^{(n)}|dy + C|\psi_k^{(n)}(0,t)|^3 \\ \le\ & C(m,M_0)(1+\|\psi_k^{(n)}(t)\|^2) + \frac{2v_bu_b}{v_b-\bar v}[\psi_k^{(n)}(0,t)]^2. \end{aligned} \tag{3.19}$$

Thus, we have $\|\psi_k^{(n)}(t)\|^2 \le (2M)^2$ due to $t_0 = t_0(m, M_0) \ll 1$ and

$$\int_0^{t_0} \|\psi_{ky}^{(n)}\|^2 dt \le \frac{v_+}{\mu}(2M)^2. \tag{3.20}$$

Multiplying (3.16) by $-\psi_{kyy}^{(n)}$ and integrating it over $\mathbb{R}_+$, one has, if δ_0 is suitably small,

$$\begin{aligned} &\left(\|\psi_{ky}^{(n)}\|^2 + \frac{2v_bu_b}{v_b-\bar v}[\psi_k^{(n)}(0,t)]^2 + \frac{2v_b}{3(v_b-\bar v)}[\psi_k^{(n)}(0,t)]^3\right)_t \\ &+\frac{\mu}{2v_+}\|\psi_{kyy}^{(n)}(t)\|^2 \le C(m,M_0) + \|\psi_{ky}^{(n)}\|^2. \end{aligned} \tag{3.21}$$

The integration of (3.21) over $(0,t)$ gives

$$\begin{aligned} &\|\psi_{ky}^{(n)}\|^2 + \frac{2v_bu_b}{v_b-\bar v}[\psi_k^{(n)}(0,t)]^2 + \frac{2v_b}{3(v_b-\bar v)}[\psi_k^{(n)}(0,t)]^3 \\ &+ \frac{\mu}{2v_+}\int_0^t \|\psi_{kyy}^{(n)}(\tau)\|^2 d\tau \\ \le\ & \|\psi_{0k}\|_1^2 + \frac{2v_bu_b}{v_b-\bar v}[\psi_{0k}(0)]^2 + \frac{2v_b}{3(v_b-\bar v)}|\psi_{0k}(0)|^3 \\ &+ C(m,M_0)t + \int_0^t \|\psi_{ky}^{(n)}\|^2 d\tau \\ \le\ & (2M)^2\left(1+\frac{2v_+}{\mu}+\frac{2v_bu_b}{v_b-\bar v}\right) \le (2M)^2\left(1+\frac{4v_b}{\mu}+\frac{2v_b}{v_b-\bar v}\right). \end{aligned} \tag{3.22}$$

Thus, we have

$$\|\psi_{ky}^{(n)}\|^2 \le \left(3\left(1+2\sqrt{\frac{v_b}{\mu}}+\sqrt{\frac{2v_b}{v_b-\bar v}}\right)M\right)^2 = (bM)^2 \tag{3.23}$$

and

$$\int_0^{t_0} \|\psi_{kyy}^{(n)}\|^2 dt \le \frac{2v_+}{\mu}(bM)^2. \tag{3.24}$$

On the other hand, a direct estimate on (3.18) together with (3.20), (3.23), and (3.24) gives

$$\|\phi_k^{(n)}(t)\|_1^2 \le (3M)^2 \tag{3.25}$$

and $\inf_{R_+\times[0,T]}(V+\phi_k^{(n)})(y,t) \geq \frac{1}{2}m$.

Therefore, $(\phi_k^{(n)}, \psi_k^{(n)}) \in X_{\frac{1}{2}m,bM}(0,t_0)$. From (3.17), it is easy to see $\gamma_k^{(n)}(t) \in C^1([0,t_0])$. By a standard way, $(\phi_k^{(n)}, \psi_k^{(n)})$ is a Cauchy sequence in $C(0,t_0;H^2)$. Thus we have a solution $(\phi_k(y,t), \psi_k(y,t), \gamma_k(t))$ by letting n tend to infinity, where $\gamma_k(t) = \int_0^t \frac{1}{\bar{v}-v_b}\psi_k(0,\tau)d\tau$. In the same way, letting $k \to \infty$, we obtain the desired unique-local solution $(\phi(y,t), \psi(y,t), \gamma(t))$ to (3.6), which satisfies $(\phi,\psi) \in X_{\frac{1}{2}m,bM}(0,T_0)$ and $\gamma(t) \in C^1([0,T_0])$ (taking T_0 smaller than t_0 if necessary).

4. Stability of the traveling wave solution. This section is devoted to the stability of the traveling wave solution. Our stability theorem is as follows.

THEOREM 4.1. *For any given $0 < \bar{v} < v_b < +\infty$, there exist positive constants δ_1 and C_1. If $(v_+,u_+) \in TW_{\bar{v}}(v_b,0)$, $v_+ - v_b = \delta \leq \delta_1$, $v_0 - V \in H_0^1$, $u_0 - U \in H^1$, and $\|v_0 - V, u_0 - U\|_1 \leq C_1\delta$, then there exists a global solution $(v(x,t), u(x,t), x(t))$ to* (1.12) *satisfying*

$$(v-V, u-U)(x,t) \in C(0,+\infty; H^1(x(t),+\infty)),$$

$$(v-V)_x(x,t) \in L^2(0,+\infty; L^2(x(t),+\infty)),$$

$$(u-U)_x(x,t) \in L^2(0,+\infty; H^1(x(t),+\infty)), \quad x(t) \in C^1([0,+\infty)),$$

$$x'(t) - \bar{s} \in L^2(0,+\infty),$$

and

$$\sup_{x\geq x(t)} |(v,u)(x,t) - (V,U)(x-x(t))| \to 0 \quad \textit{as} \quad t \to +\infty.$$

Theorem 4.1 is derived directly from local existence (Proposition 3.1) and the following a priori estimates.

PROPOSITION 4.2. *For any given $\bar{v}$ and v_b, there exist positive constants δ_1 and C_0. If $v_+ - v_b = \delta \leq \delta_1 (\leq \delta_0)$, $(\phi,\psi) \in X_{\frac{1}{4}v_b,bM}(0,T)$ with $M \leq \frac{C_0}{b}\delta$ is a solution to* (3.6) *for some positive constant T, then*

$$\begin{aligned} &\|(\phi,\psi)(t)\|_1^2 + \delta^{\frac{1}{2}}\int_0^t \psi^2(0,\tau)d\tau + \delta^{\frac{1}{2}}\int_0^t\int_0^\infty \phi^2 V'(y)dyd\tau \\ &+ \int_0^t \|\phi_y(\tau)\|^2 d\tau + \int_0^t \|\psi_y(\tau)\|_1^2 d\tau \leq C\|(\phi_0,\psi_0)\|_1^2, \end{aligned} \tag{4.1}$$

where $C > 1$ is a constant.

Proof. By virtue of the previous section, (ϕ,ψ) satisfies

$$
\begin{cases}
\phi_t-(\bar{s}+\gamma'(t))\phi_y-\psi_y=\gamma'(t)V'(y), & y>0,\quad t>0,\\
\psi_t-(\bar{s}+\gamma'(t))\psi_y+(p(V+\phi)-p(V))_y,\\
-\mu\big(\dfrac{\psi_y}{V+\phi}+\dfrac{\bar{s}V'(y)\phi}{V(V+\phi)}\big)_y=-\bar{s}\gamma'(t)V'(y), & y>0,\quad t>0,\\
\phi(0,t)=0,\\
\mu\psi_y(0,t)=\dfrac{2v_bu_b}{v_b-\bar{v}}\psi(0,t)+\dfrac{v_b}{v_b-\bar{v}}\psi^2(0,t),\\
\gamma'(t)=\dfrac{1}{\bar{v}-v_b}\psi(0,t),\quad \gamma(0)=0,\\
(\phi,\psi)|_{(+\infty,t)}=(0,0),\\
(\phi,\psi)(y,0)=(\phi_0,\psi_0)=(v_0-V,u_0-U)(y).
\end{cases}
\tag{4.2}
$$

Let

$$
\Phi(v,V)=p(V)\phi-\int_V^{V+\phi}p(s)ds.
$$

Multiplying the first equation of (4.2) by $(p(V)-p(V+\phi))$ and the second one by ψ, we have

$$
\begin{aligned}
&\left\{\left[\frac{1}{2}\psi^2+\Phi(v,V)\right]\right\}_t-(\bar{s}+\gamma')\Delta\phi V_y\\
&+\left\{(p(V+\phi)-p(V))\psi-(\bar{s}+\gamma')\left(\Phi(v,V)+\frac{1}{2}\psi^2\right)\right.\\
&\left.-\mu\left(\frac{\psi_y}{V+\phi}+\frac{\bar{s}V_y\phi}{V(V+\phi)}\right)\psi\right\}_y+\mu\frac{\psi_y^2}{V+\phi}\\
&+\frac{\mu\bar{s}V_y\psi_y\phi}{V(V+\phi)}+(p(V+\phi)-p(V))\gamma'V_y+\bar{s}\gamma'V_y\psi=0,
\end{aligned}
\tag{4.3}
$$

where $\Delta\phi=p(V+\phi)-p(V)-p'(V)\phi$ satisfies $c\phi^2\le\Delta\phi\le C\phi^2$.

Let $E=\int_0^\infty\frac{1}{2}\psi^2+\Phi(v,V)dy$. The integration of (4.3) over $R_+\times[0,T]$ gives

$$
\begin{aligned}
&E(t)-E(0)+\mu\int_0^t\int_0^\infty\frac{\psi_y^2}{V+\phi}dyd\tau-\int_0^t\int_0^\infty(\bar{s}+\gamma')\Delta\phi V_ydyd\tau\\
&+\int_0^t\left[\frac{1}{2}(\bar{s}+\gamma')\psi(0,\tau)+\mu\frac{\psi_y(0,\tau)}{v_b}\right]\psi(0,\tau)d\tau+\int_0^t\int_0^\infty\frac{\mu\bar{s}V_y\psi_y\phi}{V(V+\phi)}dyd\tau\\
&+\int_0^t\int_0^\infty(p(V+\phi)-p(V))\gamma'V_ydyd\tau+\int_0^t\int_0^\infty\bar{s}\gamma'V_y\psi dyd\tau=0.
\end{aligned}
\tag{4.4}
$$

Since $\bar{s}=O(\delta^{\frac{1}{2}})$ is negative and $|\gamma'|=O(\delta)$, we have

$$
-(\bar{s}+\gamma')\Delta\phi V_y\ge c\delta^{\frac{1}{2}}\phi^2V_y.
\tag{4.5}
$$

On the other hand, the boundary conditions of (4.2) give

$$\mu\frac{\psi_y(0,\tau)}{v_b}\psi(0,\tau)=\frac{2u_b}{v_b-\bar v}\psi^2(0,\tau)+\frac{1}{v_b-\bar v}\psi^3(0,\tau)\ge-\frac{3}{2}\bar s\psi^2(0,\tau), \tag{4.6}$$

which implies

$$\left[\frac{1}{2}(\bar s+\gamma')\psi(0,\tau)+\mu\frac{\psi_y(0,\tau)}{v_b}\right]\psi(0,\tau)\ge-\frac{1}{2}\bar s\psi^2(0,\tau)\ge c\delta^{\frac12}\psi^2(0,\tau). \tag{4.7}$$

Thus, from (4.4)–(4.7), we have

$$\begin{aligned}&E(t)-E(0)+\mu\int_0^t\int_0^\infty\frac{\psi_y^2}{V+\phi}dyd\tau+c\delta^{\frac12}\int_0^t\int_0^\infty\phi^2V_ydyd\tau\\&+c\delta^{\frac12}\int_0^t\psi^2(0,\tau)d\tau=-\int_0^t\int_0^\infty\frac{\mu\bar sV_y\psi_y\phi}{V(V+\phi)}dyd\tau\\&-\int_0^t\int_0^\infty(p(V+\phi)-p(V))\gamma'V_ydyd\tau-\int_0^t\int_0^\infty\bar s\gamma'V_y\psi dyd\tau.\end{aligned} \tag{4.8}$$

We now estimate each term on the right-hand side of (4.8). From Lemma 2.1 and

$$c\phi^2\le\Phi(v,V)\le C\phi^2, \tag{4.9}$$

we have

$$\frac{\mu\bar sV_y\psi_y\phi}{V(V+\phi)}\le\lambda\psi_y^2+C\bar s^2\phi^2V_y^2\le\lambda\psi_y^2+C\delta^{\frac32}\phi^2V_y, \tag{4.10}$$

$$\begin{aligned}&\int_0^\infty|(p(V+\phi)-p(V))\gamma'V_y|dy\\&\le C\int_0^\infty|\gamma'V_y\phi|dy\le C|\gamma'|\|\phi_y\|\int_0^\infty y^{\frac12}V_ydy\\&\le C\delta^{\frac54}|\gamma'|\|\phi_y\|\le\lambda\delta^{\frac12}|\gamma'|^2+C\delta^2\|\phi_y\|^2,\end{aligned} \tag{4.11}$$

and

$$\begin{aligned}\int_0^\infty\bar s\gamma'V_y\psi dy&=-\bar s\gamma'\int_0^\infty(V-v_+)\psi_ydy-\bar s(v_b-v_+)\gamma'\psi(0,\tau)\\&\le\lambda\|\psi_y\|^2+C\delta^{\frac32}\psi^2(0,\tau),\end{aligned} \tag{4.12}$$

where we have used the fact that

$$|\phi(y,\tau)|=\left|\int_0^y\phi_ydy\right|\le y^{\frac12}\|\phi_y\|,$$

due to [8] and λ is a suitably small positive constant.

Combining (4.8)–(4.12), we get the basic lemma.

LEMMA 4.3. *Let the conditions of Proposition* 4.2 *hold. Then*

$$\begin{aligned}&\|(\phi,\psi)(t)\|^2+\int_0^t\|\psi_y(\tau)\|^2d\tau+\delta^{\frac12}\int_0^t\psi^2(0,\tau)d\tau+\delta^{\frac12}\int_0^t\int_0^{+\infty}V_y\phi^2dyd\tau\\&\le C\|(\phi,\psi)(0)\|^2+C\delta^2\int_0^t\|\phi_y(\tau)\|^2d\tau.\end{aligned}\tag{4.13}$$

Following [17], we adapt a new variable,

$$\tilde v=\frac{v}{V},\tag{4.14}$$

to estimate $\|\phi_y\|$. Then $(4.2)_2$ is rewritten as

$$\begin{aligned}&\left(\mu\frac{\tilde v_y}{\tilde v}-\psi\right)_t-(\bar s+\gamma'(t))\left(\mu\frac{\tilde v_y}{\tilde v}-\psi\right)_y-p'(v)V\tilde v_y+(p'(V)-p'(v)\tilde v)V_y\\&=\mu\left(\frac{V_y}{V}\right)_y\gamma'(t)-\bar s\gamma'(t)V_y.\end{aligned}\tag{4.15}$$

Thus, we have the following lemma.

LEMMA 4.4. *It holds that*

$$\begin{aligned}&\|\frac{\tilde v_y}{\tilde v}(t)\|^2+\int_0^t\int_0^{+\infty}\frac{\tilde v_y^2}{\tilde v^{\gamma+2}}dyd\tau\\&\le C(\|\phi_0\|_1^2+\|\psi_0\|^2).\end{aligned}\tag{4.16}$$

Proof. Multiplying (4.15) by $\frac{\tilde v_y}{\tilde v}$, one gets

$$\begin{aligned}&\left\{\frac{\mu}{2}\left(\frac{\tilde v_y}{\tilde v}\right)^2-\psi\left(\frac{\tilde v_y}{\tilde v}\right)\right\}_t-p'(v)V\frac{\tilde v_y^2}{\tilde v}\\&+\left\{\psi\frac{\tilde v_t}{\tilde v}-\frac{\mu(\bar s+\gamma'(t))}{2}\left(\frac{\tilde v_y}{\tilde v}\right)^2\right\}_y\\&=\frac{\psi_y^2}{v}+\frac{\bar s\phi\psi_yV_y}{vV}+(p'(V)-p'(v)\tilde v)V_y\frac{\tilde v_y}{\tilde v}\\&+\left(\mu\left(\frac{V_y}{V}\right)_y\gamma'-\bar s\gamma'(t)V_y\right)\frac{\tilde v_y}{\tilde v}+\gamma'(t)\psi_y\frac{V_y}{V}.\end{aligned}\tag{4.17}$$

We compute

$$\left|\left(\frac{\tilde v_y}{\tilde v}\right)(0,t)\right|=\left|\frac{\phi_y(0,t)}{v_b}\right|=\left|\frac{\gamma'(t)V'(0)+\psi_y(0,t)}{v_b(\bar s+\gamma'(t))}\right|\le C|\psi(0,t)|,\tag{4.18}$$

$$\left|(p'(V)-p'(v)\tilde v)V_y\frac{\tilde v_y}{\tilde v}\right|\le\lambda\tilde v_y^2+CV_y^2\phi^2,\tag{4.19}$$

and

$$\begin{aligned}
&\int_0^\infty \left| \left(\frac{V_y}{V}\right)_y \gamma'(\tau)\frac{\tilde{v}_y}{\tilde{v}} \right| \\
&\le \int_0^\infty \lambda \tilde{v}_y^2 dy + C\int_0^\infty \left| \left(\frac{V_y}{V}\right)_y \right|^2 dy|\gamma'(\tau)|^2 \\
&\le \int_0^\infty \lambda \tilde{v}_y^2 dy + C\delta^{\frac{1}{2}}|\gamma'(\tau)|^2,
\end{aligned} \tag{4.20}$$

where we have used $V'(0) = O(\delta^{\frac{1}{2}})$ and $\psi_y(0,t) = O(\delta^{\frac{1}{2}})|\psi(0,t)|$. It is noted that

$$c_1\phi_y^2 - c_2V_y^2\phi^2 \le \left|\frac{\tilde{v}_y}{\tilde{v}}\right|^2 \le C_1\phi_y^2 + C_2V_y^2\phi^2,$$

where $c_i, C_i, i = 1, 2$, are positive constants which only depend on $\bar{v}$ and v_b. Integrating (4.17) over $(0, +\infty) \times (0, t)$ and using Lemma 4.3, we get Lemma 4.4.

LEMMA 4.5. *It holds that*

$$\|\psi_y(t)\|^2 + \int_0^t \|\psi_{yy}(\tau)\|^2 d\tau \le C\|\phi_0, \psi_0\|_1^2. \tag{4.21}$$

Proof. Multiplying $(4.2)_2$ by $-\psi_{yy}$, we have

$$\begin{aligned}
&\left(\frac{\psi_y}{2}\right)_t^2 + \left(-\psi_t\psi_y + \frac{\bar{s}+\gamma'(t)}{2}\psi_y^2\right)_y + \mu\frac{\psi_{yy}^2}{v} \\
&= \left\{-\mu\tfrac{\psi_y(V_y+\phi_y)}{(V+\phi)^2} + \mu\left(\tfrac{\bar{s}V_y\phi}{(V+\phi)V}\right)_y - (p(V+\phi)-p(V))_y\right\}(-\psi_{yy}) \\
&\quad + \bar{s}\gamma'(t)V_y\psi_{yy}.
\end{aligned} \tag{4.22}$$

Integrating (4.22) over $(0, +\infty) \times (0, t)$, one has

$$\begin{aligned}
&\frac{1}{2}\|\psi_y(t)\|^2 + \int_0^t \psi_t(0,\tau)\psi_y(0,\tau)d\tau - \int_0^t \frac{\bar{s}+\gamma'(t)}{2}\psi_y^2(0,\tau)d\tau \\
&\quad + \frac{\mu}{2v_+}\int_0^t \|\psi_{yy}(\tau)\|^2 d\tau \\
&\le C(\|\phi_0\|_1^2 + \|\psi_0\|_1^2) + C\int_0^t \|\phi_y(\tau)\|^2 + \|\psi_y(\tau)\|^2 d\tau \\
&\quad + C\int_0^t\int_0^\infty \bar{s}^2V_{yy}^2\phi^2 dyd\tau.
\end{aligned} \tag{4.23}$$

By the boundary conditions of (4.2), we have

$$\psi_t(0,t)\psi_y(0,t) = \frac{v_b}{v_b - \bar{v}}\left\{u_b\psi^2(0,t) + \frac{1}{3}\psi^3(0,t)\right\}_t. \tag{4.24}$$

On the other hand,

$$\int_0^t\int_0^\infty \bar{s}^2V_{yy}^2\phi^2 dyd\tau \le C\delta\int_0^t\int_0^\infty |V_{yy}|ydy \cdot \|\phi_y\|^2 d\tau \le C\delta^2\int_0^t \|\phi_y\|^2 d\tau.$$

Thus, combining (4.22)–(4.24) and the fact that $\bar{s} < 0$, we get Lemma 4.5.

Proposition 4.2 is obtained at once from Lemmas 4.3–4.5.

Proof of Theorem 4.1. Let δ_1, C_0, C, and M_0 be the constants in Propositions 3.1 and 4.2. We assume $C_0\delta_1 \le M_0$. This is possible because we can choose δ_1 smaller if necessary. We now let $v_+ - v_b = \delta \le \delta_1$ and the initial data $\|(\phi_0, \psi_0)\|_1 \le M \le \frac{C_0\delta}{\sqrt{C}b} \le M_0/b$. Then by Proposition 3.1 there exists a positive time $t_0 = t_0(v_b, M_0)$ such that there is a unique local solution (ϕ, ψ, γ) to (3.6) satisfying

$$\phi(y,t), \psi(y,t) \in X_{\frac{v_b}{4}, bM}(0, t_0), \quad \gamma(t) \in C^1([0, t_0]).$$

It is noted that $bM \le C_0\delta$. Thus, by virtue of Proposition 4.2, we have

$$\|(\phi, \psi)(t_0)\|_1 \le \sqrt{C}\|(\phi_0, \psi_0)\|_1 \le C_0\delta/b \le M_0/b,$$

which satisfies the conditions of Proposition 3.1. Hence, from the time $t = t_0$, again using Propositions 3.1 and 4.2, we know there exists a solution (ϕ, ψ, γ) to (3.6) in the interval $[t_0, 2t_0]$ and

$$\|(\phi, \psi)(2t_0)\|_1 \le \sqrt{C}\|\phi_0, \psi_0\|_1 \le C_0\delta/b \le M_0/b.$$

Repeating the above procedure, we obtain the asymptotic stability theorem, Theorem 4.1.

5. Superposition of a traveling wave solution and a rarefaction wave. This section is devoted to the superposition of a traveling wave solution and a rarefaction wave. Assume that

$$(v_+, u_+) \in R_2 TW_{\bar{v}}(v_b, 0), \tag{5.1}$$

where $0 < \bar{v} < v_b$.

In this case, there exists $(v_*, u_*) \in TW_{\bar{v}}(v_b, 0)$ such that $(v_+, u_+) \in R_2(v_*, u_*)$, and the superposition of the traveling wave solution connecting (v_b, u_b) with (v_*, u_*) and the 2-rarefaction wave connecting (v_*, u_*) with (v_+, u_+).

We now consider the 2-rarefaction wave $(v^R, u^R)(\frac{x}{t})$ connecting (v_*, u_*) with (v_+, u_+), which is the weak solution of the *Riemann* problem

$$\begin{cases} v_t - u_x = 0, \quad (x,t) \in \Re \times (0, +\infty), \\ v_t + p(v)_x = 0, \\ (v,u)\big|_{t=0} = (v_0^R, u_0^R)(x) = \begin{cases} (v_*, u_*), & x < 0, \\ (v_+, u_+), & x > 0. \end{cases} \end{cases} \tag{5.2}$$

It is known that $(v^R, u^R)(\frac{x}{t})$ has the explicit form

$$\begin{aligned} &(v^R, u^R)\left(\frac{x}{t}\right) \\ &= \begin{cases} (v_*, u_*), & -\infty \le \dfrac{x}{t} \le \lambda_2(v_*), \\ \left(\lambda_2^{-1}\left(\dfrac{x}{t}\right), u_* - \displaystyle\int_{v_*}^{\lambda_2^{-1}(\frac{x}{t})} \lambda_2(s)ds\right), & \lambda_2(v_*) \le \dfrac{x}{t} \le \lambda_2(v_+), \\ (v_+, u_+), & \lambda_2(v_+) \le \dfrac{x}{t} \le +\infty, \end{cases} \end{aligned} \tag{5.3}$$

where $\lambda_2(v) = \sqrt{-p'(v)}$.

To study the large time behavior of solutions to (1.12), it is necessary to construct a smooth approximate rarefaction wave $(\tilde{V}, \tilde{U})(x,t)$ of $(v^R, u^R)(\frac{x}{t})$ in $\Re \times (0, +\infty)$ and its restriction $(V^R, U^R)(x,t) := (\tilde{V}, \tilde{U})(x,t)\big|_{x \geq x(t)}$. For this reason, we investigate the *Riemann* problem on $\Re \times (0, +\infty)$ of the *Burgers* equation

$$\begin{cases} w_t^R + w^R w_x^R = 0, \quad (x,t) \in \Re \times \Re_+, \\ w^R(x,0) = w_0^R(x) = \begin{cases} w_- = \lambda_2(v_*), & x < 0, \\ w_+ = \lambda_2(v_+), & x > 0. \end{cases} \end{cases} \tag{5.4}$$

Here $0 < w_- < w_+$. The weak solution of (5.4) with the entropy condition is a rarefaction wave $w^R(\frac{x}{t})$ connecting w_- and w_+,

$$w^R\left(\frac{x}{t}\right) = \begin{cases} w_-, & x \leq w_- t, \\ \dfrac{x}{t}, & w_- t < x < w_+ t, \\ w_+. & w_+ t \leq x. \end{cases} \tag{5.5}$$

We now approximate $w^R(\frac{x}{t})$ by

$$\begin{cases} w_t + w w_x = 0, \quad (x,t) \in \Re \times (0, +\infty), \\ w\big|_{t=0} = w_0(x) \\ \qquad = \begin{cases} w_-, & x < 0, \\ w_- + \tilde{w}\kappa_q \displaystyle\int_0^{\varepsilon x} z^q e^{-z} dz, & x \geq 0. \end{cases} \end{cases} \tag{5.6}$$

Here $\tilde{w} = w_+ - w_-$, κ_q is a constant such that $\kappa_q \int_0^{+\infty} z^q e^{-z} dz = 1$ for large constant $q \geq 8$, and ε is a positive constant determined later. We have the following lemma.

LEMMA 5.1. *Let $0 < w_- < w_+$. Then the problem* (5.6) *has a unique smooth solution $w(x,t)$ satisfying the following:*

(i) $w_- \leq w(x,t) < w_+, w_x \geq 0$ *for* $x \geq 0, t \geq 0$.

(ii) *For any $p (1 \leq p \leq +\infty)$, there exists a constant $C_{p,q}$ such that for $t \geq 0$,*

$$\|w_x(\cdot,t)\|_{L^p} \leq C_{p,q} \min\left(\tilde{w}\varepsilon^{1-\frac{1}{p}}, \tilde{w}^{\frac{1}{p}} t^{-1+\frac{1}{p}}\right),$$
$$\|w_{xx}(\cdot,t)\|_{L^p} \leq C_{p,q} \min\left(\tilde{w}\varepsilon^{2-\frac{1}{p}}, \tilde{w}^{\frac{1}{q}}\varepsilon^{1-\frac{1}{p}+\frac{1}{q}} t^{-1+\frac{1}{q}}\right).$$

(iii) *When* $x \leq 0$, $w(x,t) - w_- = w_x(x,t) = w_{xx}(x,t) = 0$.

(iv) $\limsup_{t \to +\infty, x \in \Re} \left|w(x,t) - w^R(x,t)\right| = 0$.

Proof. Since the solution $w(x,t)$ of (5.6) has the explicit form

$$w(x,t) = w_0(x_0(x,t)), x = x_0(x,t) + w_0(x_0(x,t))t,$$

and

$$w_0'(x_0) = \begin{cases} 0, & x_0 \leq 0, \\ \tilde{w} k_q \varepsilon (\varepsilon x_0)^q e^{-\varepsilon x_0}, & x_0 > 0, \end{cases}$$

$$|w_0''(x_0)| \leq C \tilde{w}^{\frac{1}{q}} \varepsilon^{1+\frac{1}{q}} |w_0'(x_0)|^{1-\frac{1}{q}} e^{-\frac{\varepsilon x_0}{2q}},$$

it is not difficult to get Lemma 5.1 by virtue of the method of [17]. We omit the details here.

Now we define the approximate solution $(\tilde{V},\tilde{U})(x,t)$ as follows:

$$(\tilde{V},\tilde{U})(x,t)=\left(\lambda_2^{-1}\big(w(x,t)\big),u_*-\int_{v_*}^{\lambda_2^{-1}\left(w(x,t)\right)}\lambda_2(s)ds\right). \tag{5.7}$$

Setting

$$(V^R,U^R)(y,t):=(\tilde{V},\tilde{U})\big|_{x\geq x(t)},\quad y=x-x(t). \tag{5.8}$$

Since $w(x,t)$ is the smooth solution of the problem (5.6), it is easy to see

$$\begin{cases} V_t^R-(\bar{s}+\gamma'(t))V_y^R-U_y^R=0,\\ U_t^R-(\bar{s}+\gamma'(t))U_y^R+p(V^R)_y=0,\quad (y,t)\in R_+\times(0,+\infty),\\ (V^R,U^R)\big|_{y=0}=(v_*,u_*),\\ (V^R,U^R)\big|_{t=0}=\left(\lambda_2^{-1}(w_0(y)),u_*-\displaystyle\int_{v_*}^{\lambda_2^{-1}(w_0(y))}\lambda_2(s)ds\right). \end{cases} \tag{5.9}$$

Note that $|V_{yy}^R|\leq C(|w_{xx}|+|w_x|^2)$, and one has the following from Lemma 5.1.

LEMMA 5.2. *Let $\delta_2=|v_+-v_*|+|u_+-u_*|$. Then $(V^R,U^R)(y,t)$ satisfies the following if $q\geq p$:*

(i) *$U_y^R(y,t)\geq 0, |U_y^R|\leq C\varepsilon\delta_2$ for $t\geq 0, y\geq 0$.*

(ii) *For any $p(1\leq p\leq+\infty)$, there exists a constant $C_{p,q}$ such that*

$$\|V_y^R(\cdot,t)\|_{L^p(\{y\geq 0\})}\leq C_{p,q}\min\left\{\delta_2\varepsilon^{1-\frac{1}{p}},\delta_2^{\frac{1}{p}}(1+t)^{-1+\frac{1}{p}}\right\},$$
$$\|V_{yy}^R(\cdot,t)\|_{L^p(\{y\geq 0\})}\leq C_{p,q}\min\left\{\delta_2\varepsilon^{2-\frac{1}{p}},\delta_2^{\frac{1}{q}}(1+t)^{-1+\frac{1}{q}}\right\},\quad t\geq 0.$$

(iii) $(V^R,U^R)|_{y\leq -x(t)}=(v_*,u_*),\quad (V_y^R,U_y^R,V_{yy}^R,U_{yy}^R)|_{y\leq -x(t)}=0.$

(iv) $\limsup_{t\to+\infty,y\in\{y\geq 0\}}|(V^R,U^R)(y,t)-(v^R,u^R)(\frac{y+x(t)}{t})|=0.$

On the other hand, the traveling wave solution $(V_B,U_B)(y)$ connecting (v_b,u_b) with (v_*,u_*) satisfies, from Lemma 2.1,

$$\begin{cases} -\bar{s}V_y-U_y=0, & y>0,\\ -\bar{s}U_y+p(V)_y=\mu\left(\dfrac{U_y}{V}\right)_y, & y>0,\\ V(0)=v_b,U(0)=u_b,\\ \mu U_y(0)=v_b(v_b-\bar{v})\bar{s}^2,\\ (V,U)|_{(+\infty)}=(v_*,u_*), \end{cases} \tag{5.10}$$

where

$$\bar{s}=\frac{u_b}{\bar{v}-v_b}<0,\quad u_b=\frac{\bar{v}-v_b}{\bar{v}-v_*}u_*,\quad u_*=(v_*-\bar{v})^{\frac{1}{2}}(p(v_b)-p(v_*))^{\frac{1}{2}}. \tag{5.11}$$

Let

$$V(y,t) = V_B(y) + V^R(y,t) - v_*, \quad U(y,t) = U_B(y) + U^R(y,t) - u_*; \tag{5.12}$$

we set the perturbation $(\phi,\psi)(y,t)$ by $(v,u)(y,t) = (V+\phi, U+\psi)(y,t)$. Then the reformulated equation is, from (1.12), (5.9), and (5.10),

$$\begin{cases} \phi_t - (\bar{s}+\gamma'(t))\phi_y - \psi_y = \gamma'(t)V_{By}, & y>0, \quad t>0, \\ \psi_t - (\bar{s}+\gamma'(t))\psi_y + (p(V+\phi)-p(V))_y \\ -\mu\left(\dfrac{\psi_y}{V+\phi} - \dfrac{U_y\phi}{V(V+\phi)}\right)_y = -\bar{s}\gamma'(t)V_{By} + G(y), & y>0, \quad t>0, \\ \phi(0,t) = 0, \\ \mu\psi_y(0,t) = \dfrac{2v_bu_b}{v_b-\bar{v}}\psi(0,t) + \dfrac{v_b}{v_b-\bar{v}}\psi^2(0,t), \\ \gamma'(t) = \dfrac{1}{\bar{v}-v_b}\psi(0,t), \quad \gamma(0) = 0, \\ (\phi,\psi)|_{(+\infty,t)} = (0,0), \\ (\phi,\psi)(y,0) = (\phi_0,\psi_0) = (v_0-V_0, u_0-U_0)(y), \end{cases} \tag{5.13}$$

where

$$G = -(p(V)-p(V_B)-p(V^R)+p(v_*)) + \mu\left(\frac{U_y}{V} - \frac{U_{By}}{V_B}\right) = -G_1 + G_2. \tag{5.14}$$

We now derive the a priori estimates like Proposition 4.2. First we fix $\bar{v}$ and v_b. Then we choose a suitably small constant δ_0 which will be given later. We assume that $v_* - v_b = \delta \le \delta_0$ and $(\psi,\psi) \in X_{\frac{1}{4}v_b,M}(0,T)$ is a solution to (5.13) with $M \le C_0\delta^{\frac{3}{5}}$ for some positive constants T and C_0.

Multiplying $(5.13)_1$ by $p(V)-p(V+\phi)$ and $(5.13)_2$ by ψ, we have

$$\begin{aligned} &\left\{\left[\frac{1}{2}\psi^2 + \Phi(v,V)\right]\right\}_t + \Delta\phi U_y^R - (\bar{s}+\gamma')\Delta\phi V_{By} \\ &+ \left\{(p(V+\phi)-p(V))\psi - (\bar{s}+\gamma')\left(\Phi(v,V) + \frac{1}{2}\psi^2\right)\right. \\ &\left. - \mu\left(\frac{\psi_y}{V+\phi} - \frac{U_y\phi}{V(V+\phi)}\right)\psi\right\}_y + \mu\frac{\psi_y^2}{V+\phi} \\ &- \frac{\mu(U_y^R+U_{By})\psi_y\phi}{V(V+\phi)} + (p(V+\phi)-p(V))\gamma'V_{By} \\ &+ \bar{s}\gamma'V_{By}\psi - G_y\psi = 0. \end{aligned} \tag{5.15}$$

Since $p''(V) > 0$, one has

$$\Delta\phi = p(V+\phi) - p(V) - p'(V)\phi = f(v,V)\phi^2 \ge 0. \tag{5.16}$$

We regard

$$Q = \Delta\phi U_y^R - \frac{\mu U_y^R\psi_y\phi}{V(V+\phi)} + \mu\frac{\psi_y^2}{v} \tag{5.17}$$

as the quadratic equation:

$$\left(\sqrt{\mu}\frac{\psi_y}{\sqrt{v}}\right)^2 - \frac{\sqrt{\mu U_y^R}}{V\sqrt{vf(v,V)}}\cdot\sqrt{\mu}\frac{\psi_y}{\sqrt{v}}\cdot\sqrt{U_y^R f(v,V)}\phi + \left(\sqrt{U_y^R f(v,V)}\phi\right)^2. \tag{5.18}$$

By Lemma 5.2, if ε is suitably small, the discriminate of (5.18) satisfies

$$D = \frac{\mu U_y^R}{V^2 v f(v,V)} - 4 < 0. \tag{5.19}$$

It is noted that all terms including $V_B(y)$ could be treated by the same methods of the previous section; thus if δ_0 is suitably small, then we have

$$\begin{aligned}
&\|(\phi,\psi)(t)\|^2 + \int_0^t \|\psi_y(\tau)\|^2 d\tau + \delta^{\frac{1}{2}}\int_0^t \psi^2(0,\tau)d\tau \\
&+ \delta^{\frac{1}{2}}\int_0^t\int_0^{+\infty} V_{By}\phi^2 dyd\tau + \int_0^t\int_0^{+\infty} U_y^R\phi^2 dyd\tau \\
&\le C\|(\phi,\psi)(0)\|^2 + C\delta^2\int_0^t\|\phi_y(\tau)\|^2 d\tau + \int_0^t\int_0^{+\infty}|G_y||\psi|dyd\tau.
\end{aligned} \tag{5.20}$$

We now estimate the last term of (5.20). We compute

$$\begin{aligned}
|G_{1y}| &= |p'(V)(V_{By}+V_y^R) - p'(V_B)V_{By} - p'(V^R)V_y^R| \\
&\le |V_{By}||V^R - v_*| + |V_y^R||V_B - v_*|.
\end{aligned} \tag{5.21}$$

It is observed that $(V^R - v_*, V_y^R)|_{y\le -\bar{s}t-\gamma(t)} = 0$; thus we have, from Lemma 2.1,

$$\begin{aligned}
\|G_{1y}\|^2 &= \int_{-\left(\bar{s}t+\gamma(t)\right)}^{+\infty}|G_{1y}|^2(y,t)dy \\
&\le C\sup_{y\ge-\left(\bar{s}t+\gamma(t)\right)}\{|V^R(y,t)-v_*|^2|V_{By}|\}\int_{-\left(\bar{s}t+\gamma(t)\right)}^{+\infty}|V_{By}|dy \\
&\quad + C\sup_{y\ge-\left(\bar{s}t+\gamma(t)\right)}\{|V_y^R(y,t)|^2|v_*-V_B|\}\int_{-\left(\bar{s}t+\gamma(t)\right)}^{+\infty}|v_*-V_B|dy \\
&\le C\delta^{\frac{3}{2}}\delta_2^2 e^{-ct}.
\end{aligned} \tag{5.22}$$

This implies, by $\|\psi\| \le C_0\delta^{\frac{3}{5}}$,

$$\begin{aligned}
&\int_0^t\int_0^{\infty}|G_{1y}||\psi|dyd\tau \\
&\le C\int_0^t\|G_{1y}\|\|\psi\|d\tau \le C\delta^{\frac{27}{20}}.
\end{aligned} \tag{5.23}$$

On the other hand, we calculate

$$G_2 = \mu\left(\frac{U_y}{V} - \frac{U_{By}}{V_B}\right) = \mu\left(\frac{U_y^R}{V} - \frac{U_{By}(V^R - v_*)}{VV_B}\right) \tag{5.24}$$

and

$$|G_{2y}| \le C(|V^R_{yy}| + |V^R_y|^2 + |V^R_y||V_{By}| + |U_{Byy}||V^R - v_*| + |U_{By}||V^R - v_*||V_{By}|). \tag{5.25}$$

From Lemma 5.2 and (5.25), we have

$$\int_0^\infty |G_{2y}|dy \le C(\|V^R_{yy}\|_{L^1} + \delta e^{-ct} + \varepsilon^{\frac{1}{6}}(1+t)^{-\frac{5}{6}}). \tag{5.26}$$

Thus,

$$\begin{aligned}
&\int_0^t \int_0^\infty |G_{2y}||\psi|dyd\tau \\
&\le C\int_0^t \|\psi\|^{\frac{1}{2}}\|\psi_y\|^{\frac{1}{2}}(\|V^R_{yy}\|_{L^1} + \delta e^{-ct} + \varepsilon^{\frac{1}{6}}(1+t)^{-\frac{5}{6}})d\tau \\
&\le \lambda\int_0^t \|\psi_y\|^2 + C\left(\int_0^t \delta^{\frac{2}{5}}\|V^R_{yy}\|^{\frac{4}{3}}_{L^1} + \delta^{\frac{26}{15}}e^{-ct} + \varepsilon^{\frac{2}{9}}\delta^{\frac{2}{5}}(1+\tau)^{-\frac{10}{9}}d\tau\right) \\
&\le \lambda\int_0^t \|\psi_y\|^2 d\tau + C\varepsilon^{\frac{1}{6}}\delta^{\frac{2}{5}} + C\delta^{\frac{26}{15}} \\
&\le \lambda\int_0^t \|\psi_y\|^2 d\tau + C\delta^{\frac{7}{5}}
\end{aligned} \tag{5.27}$$

if $\varepsilon = O(\delta^6)$, where λ is a small positive constant. Therefore, we have the following estimate:

$$\begin{aligned}
&\|(\phi,\psi)(t)\|^2 + \int_0^t \|\psi_y(\tau)\|^2 d\tau + \delta^{\frac{1}{2}}\int_0^t \psi^2(0,\tau)d\tau \\
&+\delta^{\frac{1}{2}}\int_0^t\int_0^{+\infty} V_{By}\phi^2 dyd\tau + \int_0^t\int_0^{+\infty} U^R_y\phi^2 dyd\tau \\
&\le C\|(\phi,\psi)(0)\|^2 + C\delta^2\int_0^t \|\phi_y(\tau)\|^2 d\tau + C\delta^{\frac{27}{20}}.
\end{aligned} \tag{5.28}$$

Remark 5.1. If we choose the approximate waves (V^R, U^R) defined by (5.8) instead of the waves in [19], and let ε be suitably small, then it is not difficult to extend the stability theorem of [19] to the strong rarefaction wave by the same method.

The estimates of higher order derivatives are also obtained, though the calculations are rather tedious. Thus, we have the following theorem.

THEOREM 5.3 (the case $(v_+, u_+) \in R_2TW_{\bar v}(v_b, 0)$). *For any given $0 < \bar v < v_b < +\infty$, assume that $(v_+, u_+) \in R_2TW_{\bar v}(v_b, 0)$. Define $(V, U)(y,t)$ by* (5.12). *Then there exist positive constants δ_0 and C_0. If $v_* - v_b = \delta \le \delta_0$, $v_0(y) - V(y,0) \in H^1_0$, $u_0(y) - U(y,0) \in H^1$, and $\|v_0 - V(y,0), u_0 - U(y,0)\|_1 \le C_0\delta^{\frac{3}{5}}$, then there exists a global solution $(v(x,t), u(x,t), x(t))$ to* (1.12) *satisfying*

$$(v - V, u - U)(x,t) \in C(0,+\infty; H^1(x(t),+\infty)),$$

$$(v - V)_x(x,t) \in L^2(0,+\infty; L^2(x(t),+\infty)),$$

$$(u - U)_x(x,t) \in L^2(0,+\infty; H^1(x(t),+\infty)), \quad x(t) \in C^1([0,+\infty)),$$

$$x'(t) - \bar{s} \in L^2(0, +\infty),$$

and

$$\sup_{x \geq x(t)} |(v, u)(x, t) - (V, U)(x - x(t))| \to 0 \quad as \quad t \to +\infty.$$

Remark 5.2. Theorem 5.3 implies that it is not necessary for the 2-rarefaction wave to be weak, though the traveling wave solution is necessarily weak.

Acknowledgments. We would like to thank the referees, whose comments led to considerable improvement in this paper.

REFERENCES

[1] E. Fermi, *Thermodynamics,* Dover, New York, 1956.
[2] A. Friedman, *Partial Differential Equations of Parabolic Type,* Prentice-Hall, Englewood Cliffs, NJ, 1964.
[3] J. Goodman, *Nonlinear asymptotic stability of viscous shock profiles for conservation laws,* Arch. Ration. Mech. Anal., 95 (1986), pp. 325–344.
[4] W. Grenier, L. Neise, and H. Stocker, *Thermodynamics and Statistical Mechanics,* Springer-Verlag, New York, 1995.
[5] F.M. Huang, A. Matsumura, and X.D. Shi, *Viscous Shock Wave and boundary layer solution to an inflow problem for compressible viscous gas,* Comm. Math. Phys., to appear.
[6] A.M. Il'in and O.A. Oleinik, *Behavior of the solutions of the Cauchy problem for certain quasi linear equations for unbounded increase of time,* Trans. Amer. Math. Soc., 42 (1964), pp. 19–23.
[7] I.A. Kaliev and A.V. Kazhikhov, *Well-posedness of a gas-solid phase transition problem,* J. Math. Fluid Mech., 1 (1999), pp. 282–308.
[8] S. Kawashima and Y. Nikkuni, *Stability of stationary solutions to the half-space problem for the discrete Boltzmann equation with multiple collisions,* Kyushu J. Math., 54 (2000), pp. 233–255.
[9] S. Kawashima and S. Nishibata, *Stability of Stationary Waves for Compressible Navier-Stokes Equations in the Half Space,* in preparation.
[10] A. Kazhikhov, *On the theory of boundary value problems for equations of the one-dimensional time dependent motion of a viscous heat-conducting gas*, Comm. Math. Phys., 82 (1981/1982), pp. 37–62 (in Russian).
[11] T.P. Liu, *Nonlinear stability of shock waves for viscous conservation laws,* Mem. Amer. Math. Soc., 56 (1985).
[12] T.P. Liu, A. Matsumura, and K. Nishihara, *Behaviors of solutions for the Burgers equation with boundary corresponding to rarefaction waves,* SIAM J. Math. Anal., 29 (1998), pp. 293–308.
[13] A. Matsumura, *Inflow and outflow problems in the half space for a one-dimensional isentropic model system of compressible viscous gas*, Methods Appl. Anal., 8 (2001), pp. 645–666.
[14] A. Matsumura and M. Mei, *Convergence to travelling fronts of solutions of the p-system with viscosity in the presence of a boundary,* Arch. Ration. Mech. Anal., 146 (1999), pp. 1–22.
[15] A. Matsumura and K. Nishihara, *On the stability of traveling wave solutions of a one-dimensional model system for compressible viscous gas,* Japan J. Appl. Math., 2 (1985), pp. 17–25.
[16] A. Matsumura and K. Nishihara, *Asymptotics toward the rarefaction wave of the solutions of a one-dimensional model system for compressible viscous gas,* Japan J. Appl. Math., 3 (1986), pp. 1–13.
[17] A. Matsumura and K. Nishihara, *Global stability of the rarefaction wave of a one-dimensional model system for compressible viscous gas,* Comm. Math. Phys., 144 (1992), pp. 325–335.
[18] A. Matsumura and K. Nishihara, *Global asymptotics toward the rarefaction wave for solutions of viscous p-system with boundary effect,* Quart. Appl. Math., 58 (2000), pp. 69–83.
[19] A. Matsumura and K. Nishihara, *Large time behaviors of solutions to an inflow problem in the half space for a one-dimensional system of compressible viscous gas,* Comm. Math. Phys., 222 (2001), pp. 449–474.

[20] T. Nagasawa, *On the one-dimensional free boundary problem for the heat-conductive compressible viscous gas,* in Recent Topics in Nonlinear PDE IV, Kyoto, 1988, Lecture Notes in Numer. Appl. Anal. 10, Kinokuniya, Tokyo, 1989, pp. 83–99.

[21] S. Osher and J. Ralston, *L^1 stability of travelling waves with applications to convective porous media flow,* Comm. Pure Appl. Math., 35 (1982), pp. 737–751.

[22] D.H. Sattinger, *On the stability of waves of nonlinear parabolic systems,* Advances in Math., 22 (1976), pp. 312–355.

[23] X.D. Shi, *Asymptotic toward the rarefaction wave to an inflow problem for viscous p-system, supersonic case,* Acta Math. Appl. Sinica, to appear.

[24] A. Szepessy and Z.P. Xin, *Nonlinear stability of viscous shock waves,* Arch. Ration. Mech. Anal., 122 (1993), pp. 53–103.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1356–1379

ANALYSIS OF A SEMILINEAR PDE FOR MODELING STATIC SOLUTIONS OF JOSEPHSON JUNCTIONS*

J.-G. CAPUTO†, N. FLYTZANIS‡, A. TERSENOV§, AND E. VAVALIS§

Abstract. A semilinear elliptic partial differential equation problem that models the static (zero voltage) behavior of a Josephson window junction is considered. A priori estimates and differential properties of the solution are obtained. The existence of the solutions is shown and iterative methods for solving this problem are analyzed. Experimental numerical data that couple with the theoretical results are presented. Useful physical information is extracted from our analysis

Key words. semilinear elliptic equations, a priori estimates, Josephson window junction

AMS subject classifications. 35J25, 35Q53

PII. S0036141002303673

1. Introduction.

A Josephson junction is a weak link between two superconducting films separated by a thin oxide layer enabling the tunneling of Cooper pairs of electrons. The steady state operation under the action of an external magnetic field and bias with a constant external current is described by a semilinear elliptic partial differential equation (PDE) with a sinusoidal nonlinearity which arises from the Josephson tunneling current. The quantity that completely describes the electromagnetic properties of such a device is the difference $\phi(x, y)$ of the phases of the superconducting order parameters in the two films. The response of the junction to an external current and magnetic field depends crucially on the ratio of the junction dimensions $\mathcal{L}$ (length) and $\mathcal{W}$ (width) to the characteristic length of the problem, the Josephson penetration depth λ_j. Short junctions for which $\mathcal{L}, \mathcal{W} < \lambda_j$ are widely used in the static case (zero voltage) for magnetic field detection. When ϕ becomes time dependent the governing equation is of hyperbolic type, and such small junctions are used for voltage standard, while long junctions ($\mathcal{L} > \lambda_j > \mathcal{W}$) are very high frequency oscillators (> 100 GHz) used in astrophysical measurements. An in-depth presentation of the physics and the technological applications of Josephson junctions can be found in [2].

The main difficulty of the resonant fluxon operation of a long junction is its low energy output compounded by a strong impedance mismatch at the boundaries. The coupling of the Josephson junction to a cavity in the so-called window design allows a better impedance matching [6, 4]. It is also interesting for tailoring the static or zero voltage behavior of the device for specific purposes [10] like increasing the maximum allowed bias current in the absence of magnetic field. An extension of this model to inhomogeneous critical current density can be relevant for high T_c superconducting materials with grain boundaries [7]. Finally static solutions can be considered as fixed points and play an important role in computing the solutions of the associated

*Received by the editors March 13, 2002; accepted for publication (in revised form) July 24, 2002; published electronically May 12, 2003. This work was supported in part by a French–Greek collaboration agreement and PENED grant 2028.
http://www.siam.org/journals/sima/34-6/30367.html
†Laboratoire de Mathématiques, Institut de Sciences Appliquées, B.P. 8, 76131 Mont-Saint-Aignan cedex, France.
‡Physics Department, University of Crete, 71409 Heraklion, Greece.
§Mathematics Department, University of Crete, 71409 Heraklion, Greece (tersenov@math.uoc.gr, mav@math.uoc.gr).

hyperbolic time-dependent problem. In [4] we proposed a semilinear PDE problem which accurately and effectively modeled the static behavior of a window Josephson junction. This model enabled us to predict specific effects depending on the size of the cavity, such as the rescaling of λ_j and the increase of the maximum current for zero magnetic field [5].

If an annular geometry is considered, the periodic boundary conditions are appropriate and in this case an exhaustive classification and stability analysis of the solutions has already been carried out in one and two space dimensions including time [8]. For this geometry one is limited only to solutions with integer number of fluxons. Note, however, that this eliminates many of the interesting physical solutions that arise due to the finite size and the possibility of continuously introducing flux from the boundaries as we vary the magnetic field H.

The static two-dimensional Josephson junction problem was solved numerically by Barone et al. [1] only in the homogeneous junction case by introducing a damping term. This transforms the equation into a semilinear diffusion equation which can be discretized using explicit finite differences. A careful choice of the initial condition can lead to stable static solutions, but this cannot be guaranteed in general. In particular the junction with inhomogeneous properties requires special care. In all cases the multiplicity of solutions makes the choice of initial conditions very important, so that we need to address the static problem directly. Notice also that both the proof of the existence of a solution and some regularity estimates can be obtained easily in this time-dependent case but that these results cannot be extended to the static limit, which turns out to be a more difficult problem.

The derivation of this PDE model together with preliminary numerical experiments was presented in [4] and is briefly discussed in section 2, where comments on several mathematical peculiarities inherent in our problem are also included. In particular the periodic nonlinear right-hand side and the Neumann boundary conditions lead to an obvious nonuniqueness of the solution. Note also that the coefficients of the operator are nonsmooth. Using the *additional variable method* proposed in [11] and [17, 18], we first obtain a priori estimates on the gradient of the solution that are of physical interest. We then prove, under certain assumptions, the existence and uniqueness of the solution and the convergence of a fixed point linearization method. The study of the stability of the solutions is under way and will not be considered here. In particular we obtain in section 3 a priori estimates of the gradient of the solution of ϕ and show that the gradients are Hölder continuous functions. Based on these estimates we prove in section 4 the existence of the solution and show that a generalized second derivative exists in L_2. Assuming that the domain is narrow enough, we show, in the case of zero Neumann boundary conditions in one direction, that the solution does not depend on the associated variable. In section 5 we obtain additional estimates for the solution and its first derivatives only in terms of the external current and the magnetic field applied to it. Furthermore, assuming that the solution is in a given interval, we improve our a priori estimates. In section 6 we prove the convergence of an iterative method for linearizing the semilinear PDE problem. Numerical results that couple with our theoretical results are presented in section 7, which also discusses their physical relevance. Our conclusions are given in section 8.

2. The mathematical Josephson window junction model. Figure 2.1 shows a window junction for the case where the window Ω_j is a rectangle of size $\ell \times w$ centered in Ω. The spatial variation of the difference ϕ of the superconducting phases in both superconductors is modeled accurately in the case where the surface inductances

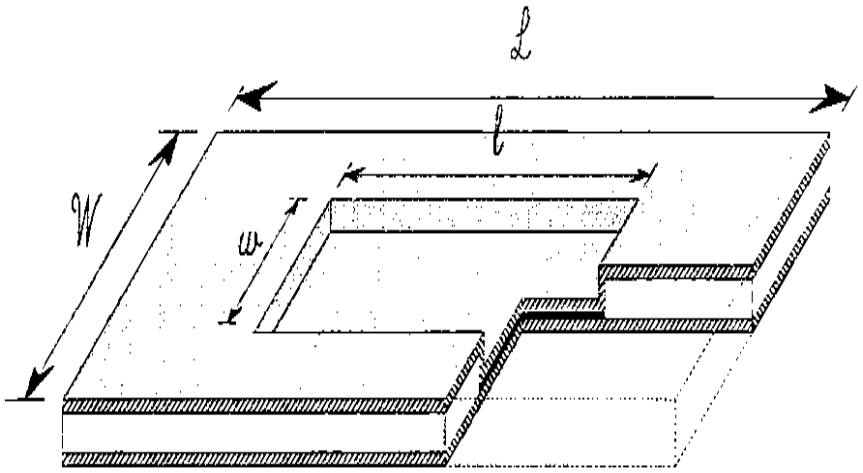


FIG. 2.1. *A window Josephson junction.*

are equal in the junction and the idle region by the equation

$$\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}=\mathcal{I}_j(x,y)\sin(\phi) \quad \text{in} \quad \Omega\equiv\left(-\frac{\mathcal{L}}{2},\frac{\mathcal{L}}{2}\right)\times\left(-\frac{\mathcal{W}}{2},\frac{\mathcal{W}}{2}\right), \tag{2.1}$$

coupled with the boundary conditions

$$\frac{\partial\phi}{\partial x}\bigg|_{x=-\frac{\mathcal{L}}{2}}=H-\alpha_1,\quad \frac{\partial\phi}{\partial x}\bigg|_{x=\frac{\mathcal{L}}{2}}=H+\alpha_2,\quad \frac{\partial\phi}{\partial y}\bigg|_{y=-\frac{\mathcal{W}}{2}}=-\delta_1,\quad \frac{\partial\phi}{\partial y}\bigg|_{y=\frac{\mathcal{W}}{2}}=\delta_2, \tag{2.2}$$

where all lengths have been normalized by λ_j. Physically $\mathcal{I}_j$ in (2.1) is the indicator function of the domain Ω_j and is discontinuous. Although in the derivation of the results that will follow we have assumed that $\mathcal{I}_j$ is continuous, we will see that all our results are independent of the smoothness of $\mathcal{I}_j$.

The model given in (2.1) can be made more realistic by including the difference in the surface inductances in the superconducting and junction regions, which leads to the equation

$$\frac{\partial}{\partial x}\left(\frac{1}{\tilde{L}(x,y)}\frac{\partial\phi}{\partial x}\right)+\frac{\partial}{\partial y}\left(\frac{1}{\tilde{L}(x,y)}\frac{\partial\phi}{\partial y}\right)=\mathcal{I}_j(x,y)\sin(\phi),$$

where $\tilde{L}$ is the normalized surface inductance. We believe that the analysis presented below can be extended to cover the case where $\tilde{L}(x,y)$ is strictly positive and differentiable.

The boundness of the right-hand side of (2.1) determines the maximum allowed values for α_1, α_2, δ_1, and δ_2. To see this, integrate both sides of (2.1) and use Green's theorem to obtain

$$\int_{\Omega_j}\sin\phi dxdy=\int_{\Omega}\nabla(\nabla\phi)dxdy=\int_{\partial\Omega}\frac{\partial\phi}{\partial n}ds=(\alpha_1+\alpha_2)\mathcal{W}+(\delta_1+\delta_2)\mathcal{L},$$

from which we have that

$$|(\alpha_1+\alpha_2)\mathcal{W}+(\delta_1+\delta_2)\mathcal{L}|\le\mu(\Omega_j), \tag{2.3}$$

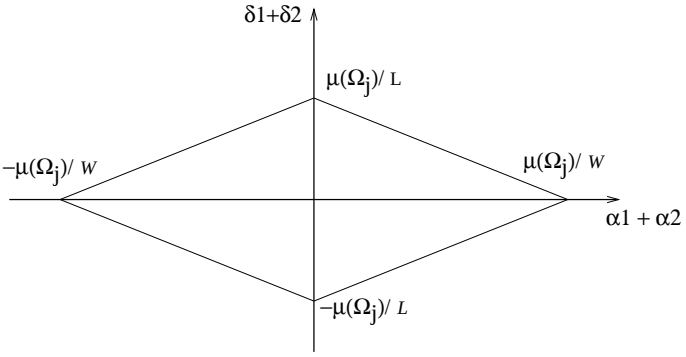


FIG. 2.2. *Allowed values for α's and δ's.*

where $\mu(\Omega_j)$ is the measure of the window domain Ω_j. From (2.3) we easily see that our PDE problem has no solutions if the α's and δ's are outside the rhombus shown in Figure 2.2.

Physically H corresponds to an external magnetic field applied in the y-direction which induces a gradient of ϕ along the x-direction. α's and δ's are current densities flowing through the device along the x- and y-directions, respectively. They can be assumed to be positive and constant. As can be seen in Figure 2.2 the sum of these currents cannot exceed the maximum critical current of the junction, which is the measure of Ω_j.

Notice also that if ϕ is a solution of the problem, then $\phi + 2k\pi, k \in Z$, is also a solution. This defines an equivalence class, so that solutions can be classified in terms of their fluxon content $n_{f\ell}$ defined by

$$n_{f\ell} \equiv \left(\sup_{\Omega} \phi - \inf_{\Omega} \phi \right) / (2\pi).$$

We also define the oscillation of $\phi(x, y)$ with respect to the variable x (and similarly for the variable y) as

$$osc_x \phi \equiv \sup_y \left(\sup_x \phi - \inf_x \phi \right). \tag{2.4}$$

Depending on the boundary conditions we can have (see [4]) a one-fluxon solution where the oscillation is between 0 and 2π, a two-fluxon solution where the oscillation is between 0 and 4π, and so on. These different solutions will have different regions of existence and different stability properties with respect to a perturbation of the boundary conditions, and as the current is increased only one will subsist. This solution gives the maximum current at zero voltage of the junction, which can be observed experimentally to indicate the quality of the junction. In the inline configuration $\alpha_1 = \alpha_2 = \alpha$, $\delta_1 = \delta_2 = 0$, in the absence of an idle region ($\Omega_j = \Omega$), Owen and Scalapino showed that the maximum current for $H = 0$ is $4\mathcal{W}$ [15]. For that they reduced the problem to one dimension and wrote the solution in terms of elliptic functions. In the same geometry but with the overlap design for which $\alpha_1 = \alpha_2 = 0$ $\delta_1 = \delta_2 = \delta$, the problem can be reduced to a one-dimensional equation only for $\mathcal{W} < 2$ [3], yielding a maximum current for $H = 0$ of $\mathcal{L} \times \mathcal{W}$. When $\mathcal{W} > 2$ the current for $H = 0$ saturates, as expected, to $4 \times \mathcal{L}$, and transverse modes are needed

for the description [3]. The presence of an idle region ($\Omega_j \neq \Omega$) has important effects on the behavior of the junction. In particular the characteristic length is larger than $\lambda_J \equiv 1$, which leads to an increase of the maximum current for $H = 0$.

An important case is when the device is symmetric with respect to the center. Then if $\delta_1 = \delta_2$, one can assume the solution to be symmetric with respect to the horizontal middle line, and if the x boundary conditions are antisymmetric, i.e., $\alpha_1 = \alpha_2$ and $H = 0$, the solution will be symmetric with respect to the vertical middle line, so that just a quarter of the device might be considered. A priori estimates have been derived for these cases also.

Notice also that the solution of the PDE problem is a minimum of the free energy functional

$$\begin{aligned} F = &\int_\Omega \left[\frac{1}{2}\left(\frac{\partial\phi}{\partial x}\right)^2 + \frac{1}{2}\left(\frac{\partial\phi}{\partial y}\right)^2 + \mathcal{I}_j\left(1-\cos\phi\right)\right] dxdy \\ &- \int_{-\frac{\mathcal{W}}{2}}^{\frac{\mathcal{W}}{2}} \left[(H-\alpha_1)\phi\left(-\frac{\mathcal{L}}{2}, y\right) - (H+\alpha_2)\phi\left(\frac{\mathcal{L}}{2}, y\right)\right] dy \\ &- \int_{-\frac{\mathcal{L}}{2}}^{\frac{\mathcal{L}}{2}} \left[-\delta_2\phi\left(x, \frac{\mathcal{W}}{2}\right) - \delta_1\phi\left(x, -\frac{\mathcal{W}}{2}\right)\right] dx. \end{aligned}$$

Due to the multiplicity of solutions there are several minima. In the simple case where the boundary conditions are nonzero only in x or y and $\Omega \equiv \Omega_j$, the y or x dependence can be neglected and the last term of the free energy can be significantly simplified.

3. A priori estimates of the gradient. The main objective of this section is to obtain estimates of the gradient of the solution that are of practical interest in either proving the existence of the solution or measuring the gradient in terms of physical quantities. Estimates of the gradient in terms of the maximum of the solution are easily obtained from well-known results [13, 9]. In this section we obtain a priori estimates for the gradient of a classical solution of the proposed PDE model only in terms of the size of the domain and the physical parameters of the problem. Note that the estimates obtained below are independent of the solution and cannot be obtained from classical results [13, 9].

We start by homogenizing the problem (2.1)–(2.2) by setting $u \equiv \phi - f$ with

$$f \equiv Hx + \frac{\alpha_1}{2\mathcal{L}}\left(x - \frac{\mathcal{L}}{2}\right)^2 + \frac{\alpha_2}{2\mathcal{L}}\left(x + \frac{\mathcal{L}}{2}\right)^2 + \frac{\delta_1}{2\mathcal{W}}\left(y - \frac{\mathcal{W}}{2}\right)^2 + \frac{\delta_2}{2\mathcal{W}}\left(y + \frac{\mathcal{W}}{2}\right)^2$$

to get from (2.1) and (2.2) that

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \mathcal{I}_j \sin(u+f) - \frac{\alpha_1+\alpha_2}{\mathcal{L}} - \frac{\delta_1+\delta_2}{\mathcal{W}} \quad \text{in } \Omega \tag{3.1}$$

and

$$\left.\frac{\partial u}{\partial x}\right|_{x=\pm\frac{\mathcal{L}}{2}} = \left.\frac{\partial u}{\partial y}\right|_{y=\pm\frac{\mathcal{W}}{2}} = 0. \tag{3.2}$$

In what follows, without explicitly stating, we assume that the indicator function $\mathcal{I}_j$ is smooth. This assumption is set only to guarantee the existence of a classical solution of the problem and does not affect the result of the lemmas since we do

not have any smoothness requirements. In practice these indicator functions are discontinuous (i.e., $\mathcal{I}_j(x,y)$ is 1 if $(x,y) \in \overline{\Omega}_j$ and 0 otherwise). To treat such $\mathcal{I}_j$ we can consider a continuously differentiable function $\mathcal{I}_j^\delta \in C^1\left(\overline{\Omega}\right)$, $0 \le \mathcal{I}_j^\delta \le 1$, such that $\mathcal{I}_j^\delta \to \mathcal{I}_j$ *weak in L_∞ for which the analysis that will follow is valid. Therefore in what follows and for simplicity in the notation we will use the symbol $\mathcal{I}_j$ instead of $\mathcal{I}_j^\delta$.

LEMMA 3.1. *For any classical solution $u(x,y)$ of the problem* (3.1)–(3.2) *we have that*

$$|u_x| \le \mathcal{L}, \qquad |u_y| \le \mathcal{W}. \tag{3.3}$$

Proof. We start by writing (3.1) at a point $(\xi, y) \in \Omega$ with $\xi \neq x$ as

$$\frac{\partial^2 u(\xi,y)}{\partial \xi^2} + \frac{\partial^2 u(\xi,y)}{\partial y^2} = \mathcal{I}_j(\xi,y)\sin\left(u(\xi,y)+f(\xi,y)\right) - \frac{\alpha_1+\alpha_2}{\mathcal{L}} - \frac{\delta_1+\delta_2}{\mathcal{W}}. \tag{3.4}$$

Now define the function $v(x,y,\xi) \equiv u(x,y) - u(\xi,y)$, for which, by subtracting (3.4) from (3.1), we have

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial \xi^2} = \mathcal{I}_j(x,y)\sin\left(u(x,y)+f(x,y)\right) - \mathcal{I}_j(\xi,y)\sin\left(u(\xi,y)+f(\xi,y)\right),$$

and thus

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial \xi^2} \ge -2. \tag{3.5}$$

Now consider the prism

$$P_1 = \left\{(x,\xi,y) : |x| < \frac{\mathcal{L}}{2}, |\xi| < \frac{\mathcal{L}}{2}, |y| < \frac{\mathcal{W}}{2}, x-\xi > 0\right\}$$

and the ordinary differential equation problem

$$h''(\tau) = -1, \quad \text{with} \quad h(0) = 0 \quad \text{and} \quad h'(\mathcal{L}) = \epsilon, \tag{3.6}$$

where ϵ is a positive constant. The solution of (3.6) is given by $h(\tau) = -\frac{\tau^2}{2} + \tau(\mathcal{L}+\epsilon)$. Define the function $\omega(x,y,\xi) \equiv v(x,y,\xi) - h(x-\xi)$ and take into account (3.5) to get that

$$\frac{\partial^2 \omega}{\partial x^2} + \frac{\partial^2 \omega}{\partial y^2} + \frac{\partial^2 \omega}{\partial \xi^2} \ge 0, \tag{3.7}$$

from which, using the strong maximum principle (see Lemma 3.5 in [9]), we conclude that ω does not achieve its maximum value in P_1 unless it is a constant function. On the boundary sector defined by $x = \frac{\mathcal{L}}{2}$, $|y| \le \frac{\mathcal{W}}{2}$, and $-\frac{\mathcal{L}}{2} \le \xi < \frac{\mathcal{L}}{2}$ we have that

$$\frac{\partial \omega(\frac{\mathcal{L}}{2},y,\xi)}{\partial x} = \frac{\partial u(\frac{\mathcal{L}}{2},y)}{\partial x} - h'\left(\frac{\mathcal{L}}{2}-\xi\right) = -h'\left(\frac{\mathcal{L}}{2}-\xi\right) < 0.$$

Since x is the outward normal to the domain, ω does not achieve its maximum on this part of the boundary. On the boundary sector defined by $\xi = -\frac{\mathcal{L}}{2}$, $|y| \le \frac{\mathcal{W}}{2}$, and $-\frac{\mathcal{L}}{2} < x \le \frac{\mathcal{L}}{2}$ we have that

$$\frac{\partial \omega(x,y,-\frac{\mathcal{L}}{2})}{\partial \xi} = -\frac{\partial u(-\frac{\mathcal{L}}{2},y)}{\partial \xi} + h'\left(x+\frac{\mathcal{L}}{2}\right) = h'\left(x+\frac{\mathcal{L}}{2}\right) > 0,$$

and since ξ is the inward normal we conclude that the maximum of ω is not achieved on this part of the boundary either. On the planes $y = \pm\frac{\mathcal{W}}{2}$, $|x| < \frac{\mathcal{L}}{2}$, and $|\xi| < \frac{\mathcal{L}}{2}$ we have that

$$\frac{\partial\omega(x,\pm\frac{\mathcal{W}}{2},\xi)}{\partial y} = \frac{\partial u(x,\pm\frac{\mathcal{W}}{2})}{\partial y} - \frac{\partial u(\xi,\pm\frac{\mathcal{W}}{2})}{\partial y} = 0,$$

and assuming that ω is not a constant function we have, from Lemma 3.4 in [9], that ω does not achieve its maximum on these boundary planes either. Therefore, the maximum is achieved at $x = \xi$, and since $\omega(x,y,\xi)|_{x=\xi} = 0$ we have that the inequality

$$u(x,y) - u(\xi,y) \le h(x-\xi)$$

holds in Ω, which, as can be easily seen, becomes an equality if ω is a constant function.

Subtracting relation (3.1) from (3.4) and applying the above analysis we obtain that

$$u(\xi,y) - u(x,y) \le h(x-\xi),$$

and hence

$$|u(x,y) - u(\xi,y)| \le h(x-\xi) \quad \text{for} \quad x > \xi.$$

Working in a similar way (or directly obtained by symmetry) for $x < \xi$ we easily see that

$$|u(x,y) - u(\xi,y)| \le h(|x-\xi|) - h(0).$$

By dividing the last relation by $|x-\xi|$ and taking the limit, we have that $\left|\frac{\partial u}{\partial x}\right| \le h'(0)$ and finally obtain the first of the following inequalities (when $\epsilon \to 0$), while the second can be obtained similarly.

$$\sup_{(x,y)\in\Omega}\left|\frac{\partial u(x,y)}{\partial x}\right| \le \mathcal{L}, \qquad \sup_{(x,y)\in\Omega}\left|\frac{\partial u(x,y)}{\partial y}\right| \le \mathcal{W}. \qquad \square$$

Remark 3.1. As a direct consequence of the above lemma, we easily get the following estimates for the gradient of the solution of the problem (2.1)–(2.2):

$$-\mathcal{L} + H + \frac{\alpha_1+\alpha_2}{\mathcal{L}}x + \frac{\alpha_2-\alpha_1}{2} \le \frac{\partial\phi}{\partial x} \le \mathcal{L} + H + \frac{\alpha_1+\alpha_2}{\mathcal{L}}x + \frac{\alpha_2-\alpha_1}{2} \tag{3.8}$$

and

$$-\mathcal{W} + \frac{\delta_1+\delta_2}{\mathcal{W}}y + \frac{\delta_2-\delta_1}{2} \le \frac{\partial\phi}{\partial y} \le \mathcal{W} + \frac{\delta_1+\delta_2}{\mathcal{W}}y + \frac{\delta_2-\delta_1}{2}. \tag{3.9}$$

It is worth noting here that as it follows from (3.8) $\phi_x > 0$ in the case of large magnetic field H. This is consistent with the physical properties of Josephson junctions [2].

Next we obtain sharper estimates by making certain assumptions, on typical junction's size, on the domain.

LEMMA 3.2. (a) *Suppose that $\mathcal{I}_j$ depends only on the variable y. If $\mathcal{L} < 2$, then for any classical solution $u(x,y)$ of the problem* (2.1)–(2.2) *we have*

$$|u_x| \le \frac{\mathcal{L}^2 f_1}{4-\mathcal{L}^2}, \tag{3.10}$$

where $f_1 = \max|f_x(x,y)|$.

(b) *Suppose that $\mathcal{I}_j$ depends only on the variable x. If $\mathcal{W} < 2$, then for any classical solution $u(x,y)$ of the problem* (2.1)–(2.2) *we have*

$$|u_y| \le \frac{\mathcal{W}^2 f_2}{4-\mathcal{W}^2}, \tag{3.11}$$

where $f_2 = \max|f_y(x,y)|$.

Proof. Arguing in the same manner as in the proof of Lemma 3.1, for $v(x,y,\xi) = u(x,y) - u(\xi,y)$ we obtain

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial \xi^2} = \mathcal{I}_j(y)\left(\sin\left(u(x,y)+f(x,y)\right) - \sin\left(u(\xi,y)+f(\xi,y)\right)\right), \tag{3.12}$$

and thus (from Lemma 3.1 we already have that $|u_x| \le \mathcal{L}$)

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial \xi^2} \ge -\mathcal{L}(x-\xi) \quad \text{for} \quad x > \xi. \tag{3.13}$$

Consider the prism

$$P_1 = \left\{(x,\xi,y) : |x| < \frac{\mathcal{L}}{2}, |\xi| < \frac{\mathcal{L}}{2}, |y| < \frac{\mathcal{W}}{2}, x-\xi > 0\right\},$$

and let $h_1(\tau)$ be a solution of the problem

$$h_1''(\tau) = -\frac{\mathcal{L}+f_1}{2}\tau, \quad h_1(0) = 0, \quad \text{and} \quad h_1'(\mathcal{L}) = \epsilon > 0. \tag{3.14}$$

Obviously for $h_1 = h_1(x-\xi)$ we have

$$\frac{\partial^2 h_1}{\partial x^2} + \frac{\partial^2 h_1}{\partial y^2} + \frac{\partial^2 h_1}{\partial \xi^2} = -(\mathcal{L}+f_1)(x-\xi). \tag{3.15}$$

Subtracting (3.15) from (3.13) for $\omega(x,y,\xi) \equiv v(x,y,\xi) - h_1(x-\xi)$ we obtain that

$$\frac{\partial^2 \omega}{\partial x^2} + \frac{\partial^2 \omega}{\partial y^2} + \frac{\partial^2 \omega}{\partial \xi^2} \ge 0. \tag{3.16}$$

Arguing analogously to the proof of Lemma 3.1 we have

$$|u_x(x,y)| \le h_1'(0) = \frac{\mathcal{L}+f_1}{4}\mathcal{L}^2 + \epsilon,$$

and passing to the limit when $\epsilon \to 0$,

$$|u_x(x,y)| \le \frac{\mathcal{L}+f_1}{4}\mathcal{L}^2 \equiv \mathcal{L}_1.$$

Returning back to (3.12) and taking into account that now $|u_x| \le \mathcal{L}_1$ we obtain

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial \xi^2} \ge -(\mathcal{L}_1 + f_1)(x-\xi) \quad \text{for} \quad x > \xi. \tag{3.17}$$

Construct the function

$$h_2''(\tau) = -\frac{\mathcal{L}_1 + f_1}{2}\tau, \quad h_2(0) = 0, \quad \text{and} \quad h_2'(\mathcal{L}) = \epsilon. \tag{3.18}$$

In a similar manner as above we conclude that

$$|u_x(x,y)| \le h_2'(0) = \frac{\mathcal{L}_1 + f_1}{4}\mathcal{L}^2 + \epsilon,$$

and letting $\epsilon \to 0$ we have

$$|u_x(x,y)| \le \frac{\mathcal{L}_1 + f_1}{4}\mathcal{L}^2 = \left(\frac{\mathcal{L}}{2}\right)^4 (\mathcal{L} + f_1) + \left(\frac{\mathcal{L}}{2}\right)^2 f_1 \equiv \mathcal{L}_2.$$

Continuing this procedure we obtain the sequence of the bounds for $|u_x|$,

$$\mathcal{L}_n = \left(\frac{\mathcal{L}}{2}\right)^{2n} \mathcal{L} + f_1 \left[\left(\frac{\mathcal{L}}{2}\right)^{2n} + \left(\frac{\mathcal{L}}{2}\right)^{2(n-1)} + \cdots + \left(\frac{\mathcal{L}}{2}\right)^4 + \left(\frac{\mathcal{L}}{2}\right)^2\right].$$

If $\mathcal{L} < 2$, then $\left(\frac{\mathcal{L}}{2}\right)^{2n} \mathcal{L} \to 0$ when $n \to \infty$, and the second term

$$f_1 \left[\left(\frac{\mathcal{L}}{2}\right)^{2n} + \cdots + \left(\frac{\mathcal{L}}{2}\right)^2\right] \to f_1 \frac{\mathcal{L}^2}{4 - \mathcal{L}^2}.$$

This concludes the proof of part (a); the proof of part (b) is similar. □

Remark 3.2. Recall that $u \equiv \phi - f$ and use the above lemma to obtain the following estimates in terms of ϕ:

$$-\frac{\mathcal{L}^2 f_1}{4-\mathcal{L}^2} + H + \frac{\alpha_1 + \alpha_2}{\mathcal{L}}x + \frac{\alpha_2 - \alpha_1}{2} \le \frac{\partial \phi}{\partial x} \le \frac{\mathcal{L}^2 f_1}{4-\mathcal{L}^2} + H + \frac{\alpha_1 + \alpha_2}{\mathcal{L}}x + \frac{\alpha_2 - \alpha_1}{2}. \tag{3.19}$$

In particular, if $f_1 = H = \alpha_1 = \alpha_2 \equiv 0$, then $\frac{\partial \phi}{\partial x} \equiv 0$.

$$-\frac{\mathcal{W}^2 f_2}{4-\mathcal{W}^2} + \frac{\delta_1 + \delta_2}{\mathcal{W}}y + \frac{\delta_2 - \delta_1}{2} \le \frac{\partial \phi}{\partial y} \le \frac{\mathcal{W}^2 f_2}{4-\mathcal{W}^2} + \frac{\delta_1 + \delta_2}{\mathcal{W}}y + \frac{\delta_2 - \delta_1}{2}. \tag{3.20}$$

In particular, if $f_2 = \delta_1 = \delta_2 \equiv 0$, then $\frac{\partial \phi}{\partial y} \equiv 0$.

Remark 3.3. If $\delta_1 = \delta_2 = 0$, $\mathcal{I}_j = \mathcal{I}_j(x)$, and $\mathcal{W} < 2$, then our problem becomes one-dimensional: $\phi(x,y) = \phi(x)$ and (2.1)–(2.2) take the form

$$\phi''(x) = \mathcal{I}_j(x)\sin\phi(x), \qquad \phi'\left(-\frac{\mathcal{L}}{2}\right) = H - \alpha_1, \qquad \phi'\left(\frac{\mathcal{L}}{2}\right) = H + \alpha_2. \tag{3.21}$$

LEMMA 3.3. *The first order derivatives of the classical solution of problem* (2.1)–(2.2) *are Hölder continuous with the Hölder coefficient and exponent depending only on* $||\frac{\partial \phi}{\partial x}||_{L^2(\Omega)}$, $||\frac{\partial \phi}{\partial y}||_{L^2(\Omega)}$, *and* $\alpha_1, \alpha_2, \delta_1, \delta_2, \mathcal{L}$, *and* $\mathcal{W}$.

The bounds on the Hölder norm of the gradient follow from Theorem 9.11 in [9].

4. Existence and uniqueness. In this section we show, under certain conditions, the existence and the uniqueness of a solution of the PDE problem (3.1)–(3.2) (and therefore of the problem (2.1)–(2.2)). We start by giving the definition of the generalized solution.

DEFINITION 4.1. *We call a function* $u(x,y) \in C^{1,\alpha}(\overline{\Omega}) \cap W_2^2(\Omega)$ *a generalized solution of the PDE problem* (3.1)–(3.2) *if it satisfies the integral identity*

$$\int_\Omega (u_{xx} + u_{yy} - \mathcal{I}_j \sin(u+f) - g)\,\psi dxdy = 0 \quad \forall \psi \in L^2(\Omega) \tag{4.1}$$

and the boundary conditions (3.2), *where* $g \equiv -\frac{\alpha_1+\alpha_2}{\mathcal{L}} - \frac{\delta_1+\delta_2}{\mathcal{W}}$.

4.1. Existence. We start by assuming that $\mathcal{I}_j$ is smooth and consider the auxiliary problem

$$\Delta v = \kappa \mathcal{I}_j \left(\sin(v+f) - \frac{1}{\mu(\Omega_j)} \int_{\Omega_j} \sin(v+f) dxdy \right) \text{ in } \Omega, \tag{4.2}$$

$$\frac{\partial v}{\partial n} = 0 \quad \text{on } \partial\Omega, \tag{4.3}$$

and

$$\frac{1}{\mu(\Omega)} \int_\Omega v dxdy = \zeta, \tag{4.4}$$

where $\kappa \in [0,1]$ and ζ is an arbitrary fixed real number. Recall that by $\mu(\Omega)$ we denote the measure of Ω. We will show that a solution $v \in C^{1,\gamma}(\bar{\Omega}) \cap C^3(\Omega)$ of the auxiliary problem (4.2)–(4.4) exists. For this we define $\psi \equiv v - \zeta$ and write the above problem in the following equivalent form:

$$\Delta \psi = \kappa \mathcal{I}_j \left[\sin(\psi + \zeta + f) - \frac{1}{\mu(\Omega_j)} \int_{\Omega_j} \sin(\psi + \zeta + f) dxdy \right] \text{ in } \Omega, \tag{4.5}$$

$$\frac{\partial \psi}{\partial n} = 0 \quad \text{on } \partial\Omega, \tag{4.6}$$

and

$$\int_\Omega \psi dxdy = 0. \tag{4.7}$$

As we easily see, the only difference between (3.1) and (4.5) is a bounded constant term on the right-hand side. Hence the estimates obtained in the lemmas in the previous section hold for (4.5)–(4.6) too. We can also observe that ψ becomes zero at least at one point in Ω so that using the estimate of the gradient (which are independent of the $\max|\psi|$) we can obtain a bound for the maximum of $|\psi|$ in the domain $\bar{\Omega}$. We are in the position now to use the Leray–Schauder theorem [9, Theorem 11.3] (see the appendix) to prove the existence of the generalized solution of problem (4.2)–(4.4).

LERAY–SCHAUDER THEOREM. *Let* T *be a compact mapping from a Banach space* $\mathcal{B}$ *to itself, and suppose there exists a constant* M *such that* $||u||_{\mathcal{B}} < M$ *for all* $u \in \mathcal{B}$ *and* $\kappa \in [0,1]$ *satisfying* $u = \kappa T u$. *Then* T *has a fixed point.*

We now address the case of a nonsmooth function $\mathcal{I}_j$.

LEMMA 4.2. *For the classical solution of the PDE problem* (4.2)–(4.4) *for* $\zeta = 1$ *the following inequality holds:*

$$\int_\Omega \left[\left(\frac{\partial^2 v}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 v}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 v}{\partial y^2} \right)^2 \right] dxdy \le 4\mu(\Omega_j). \tag{4.8}$$

Proof. We square both sides of (4.2), integrate them two times (using integration by parts for the second term on the left-hand side). Then the use of the boundary conditions easily gives the above bound. ∎

Obviously the classical solution v satisfies the integral identity (4.1), i.e.,

$$\int_\Omega \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} - \mathcal{I}_j^\delta \sin(v+f) + \frac{1}{\mu(\Omega)} \int_{\Omega_j} \sin(v+f) dxdy \right) \psi dxdy = 0 \ \forall \psi \in L^2(\Omega). \tag{4.9}$$

Taking the limit, as $\delta \to 0$ we readily obtain the existence of the generalized solution.

Let us fix arbitrarily $\zeta = \zeta_0$. For this ζ_0 we find the generalized solution of the problem (4.2)–(4.4) for $\kappa = 1$. In order to obtain the existence of the original problem we need to find boundary conditions such that

$$R(c) = c, \tag{4.10}$$

where

$$c \equiv \mu(\Omega_j) \left(\frac{\alpha_1 + \alpha_2}{\mathcal{L}} + \frac{\delta_1 + \delta_2}{\mathcal{W}} \right)$$

and

$$R(c) \equiv \int_{\Omega_j} \sin(v+f) dxdy.$$

Note that R satisfies the inequality $|R| < \mu(\Omega_j)$, and observe, assuming that R continuously depends on c, that it is impossible to have $R < c$ for c varying from $\mu(\Omega_j)$ to $-\mu(\Omega_j)$. Hence there exists such c_0 that verifies (4.10). For this c_0 the solution of the auxiliary problem (4.2)–(4.4) coincides with the solution of the original one (2.1)–(2.2). Mark that the assumption on the continuity of R is satisfied in the cases of the uniqueness of the solution of the auxiliary problem. Such uniqueness can be proved following an analysis similar to the one presented in Theorem 4.4. From the above we readily obtain the following theorem.

THEOREM 4.3. *If the solution of the problem* (4.2)–(4.4) *is unique, then for any* $\zeta \in \mathbb{R}$ *we can find values for* $H, \alpha_1, \alpha_2, \delta_1,$ *and* δ_2 *for which there exists a generalized solution* ϕ *of the problem* (2.1)–(2.2) *such that*

$$\frac{1}{\mu(\Omega)} \int_\Omega (\phi - f) dxdy = \zeta. \tag{4.11}$$

Let us note that Theorem 4.3 also holds if condition (4.11) is replaced by

$$(\phi - f)|_{(x_0, y_0)} = \zeta, \quad (x_0, y_0) \in \Omega.$$

To prove this, one has to carry out an analysis similar to the above, which is lengthy and tedious and so will not be presented here.

4.2. Uniqueness. It has been observed both numerically and experimentally [2, 3, 4] and it is intuitively expected that our PDE problem might have more than one nontrivial solution. In the case when $\mathcal{L} < 2$ and $\mathcal{W} < 2$ we can easily see from Lemma 3.2 that the only solution is $n\pi$, $n = 0, \pm 1, \pm 2, \ldots$. An extensive theoretical and experimental bifurcation analysis is under way and will be presented elsewhere. Nevertheless, as we show below, under certain conditions only one solution exists.

THEOREM 4.4. *Assume that either*

(a) $\mathcal{L} < 2$, $w < \sqrt{2}$, $\mathcal{I}_j = \mathcal{I}_j(y)$, *and* $H = \alpha_1 = \alpha_2 = 0$, *or*

(b) $\mathcal{W} < 2$, $\ell < \sqrt{2}$, $\mathcal{I}_j = \mathcal{I}_j(x)$, *and* $\delta_1 = \delta_2 = 0$.

Then the generalized solution u *of* (3.1)–(3.2) *satisfying the condition*

$$\frac{1}{\mu(\Omega)} \int_\Omega u dx dy = \zeta, \tag{4.12}$$

where ζ *is an arbitrarily given constant, is unique.*

Proof. From Lemma 3.2 it follows that $u_x \equiv 0$. Thus we have

$$u_{yy} = \mathcal{I}_j(y) \sin(u + f) - \frac{\delta_1 + \delta_2}{\mathcal{W}} \quad \text{in} \quad \Omega, \tag{4.13}$$

$$u_y \left(\pm \frac{\mathcal{W}}{2} \right) = 0, \tag{4.14}$$

and

$$\int_{-\frac{\mathcal{W}}{2}}^{\frac{\mathcal{W}}{2}} u dy = \zeta. \tag{4.15}$$

Suppose now that there exist two different solutions u and v, both satisfying condition (4.15), i.e.,

$$\int_{-\frac{\mathcal{W}}{2}}^{\frac{\mathcal{W}}{2}} u dy = \int_{-\frac{\mathcal{W}}{2}}^{\frac{\mathcal{W}}{2}} v dy. \tag{4.16}$$

This implies that u and v cross each other. Now let $\sigma \equiv u - v$ and observe that

$$\sigma_{yy} = \mathcal{I}_j(y) \left(\sin(u(y) + f(y)) - \sin(v(y) + f(y)) \right) = \mathcal{I}_j(y) \sigma \cos\theta. \tag{4.17}$$

Suppose that u and v intersect at a point y_0. Consider the two cases

$$(\alpha) \quad y_0 \notin \left[-\frac{w}{2}, \frac{w}{2} \right], \qquad (\beta) \quad y_0 \in \left[-\frac{w}{2}, \frac{w}{2} \right]. \tag{4.18}$$

In (α) consider the case $y_0 \in \left(-\frac{\mathcal{W}}{2}, \frac{w}{2}\right]$. In the interval $\left(-\frac{\mathcal{W}}{2}, y_0\right)$ we have $\sigma_{yy} = 0$ and $\sigma_y(-\frac{\mathcal{W}}{2}) = \sigma(y_0) = 0$. Hence $\sigma \equiv 0 \in \left(-\frac{\mathcal{W}}{2}, y_0\right)$. Therefore, due to the analyticity of σ in $\left(-\frac{\mathcal{W}}{2}, \frac{w}{2}\right]$ we have $\sigma \equiv 0 \in \left(-\frac{\mathcal{W}}{2}, \frac{w}{2}\right]$. Similarly we can consider the case $y_0 \in \left[\frac{w}{2}, \frac{\mathcal{W}}{2}\right)$.

Consider now the (β) case. Multiplying (4.17) by σ and integrating by parts we get

$$\int_{y_0}^{\frac{\mathcal{W}}{2}} \sigma_y^2 dy \le \int_{y_0}^{\frac{\mathcal{W}}{2}} \mathcal{I}_j \sigma^2 dy = \int_{y_0}^{\frac{w}{2}} \sigma^2 dy. \tag{4.19}$$

Applying the Poincaré inequality we obtain

$$\int_{y_0}^{\frac{w}{2}} \sigma_y^2 dy \le \int_{y_0}^{\frac{\mathcal{W}}{2}} \sigma_y^2 dy \le \int_{y_0}^{\frac{w}{2}} \sigma^2 dy \le \frac{w^2}{2} \int_{y_0}^{\frac{w}{2}} \sigma_y^2 dy. \tag{4.20}$$

Due to the assumption $w < \sqrt{2}$ we have $\int_{y_0}^{\frac{w}{2}} \sigma_y^2 dy = 0$, and therefore $\sigma_y \equiv 0$. Since $\sigma(y_0) = 0$ we have $\sigma \equiv 0$ and $u \equiv v$. □

We note that there are other cases where one might be able to show this uniqueness. For example, we have shown that if we assume that the window is such that $\ell < \sqrt{2}$ and $w < \sqrt{2}$, then the generalized solution u of (3.1)–(3.2) satisfying the condition

$$\frac{1}{\mu(\Omega)} \int_\Omega u dx dy = \zeta, \tag{4.21}$$

where ζ is an arbitrarily given constant, is unique. The proof of this statement is similar to the proof of the previous theorem. Since it is rather technical, tedious, and lengthy it is not included here.

In a manner similar to the above theorem it can be shown that there exist cases where the solution of the auxiliary problem is unique.

5. Additional estimates. In this section we obtain estimates of the gradient of the solution of the problem in some special cases that are of physical interest. As discussed in section 2 it is useful [3, 4, 5] to characterize the solutions of (2.1)–(2.2) by their oscillations, defined by (2.4). In what follows we derive estimates of the gradient of the solution as a function of its oscillations in the x- and y-directions.

LEMMA 5.1. *For any classical solution $\phi(x, y)$ of the problem* (2.1)–(2.2) *we have that*

$$\left|\frac{\partial \phi}{\partial x}\right| \le \sqrt{2 osc_x \phi + (H + \alpha_2)^2} \tag{5.1}$$

and

$$\left|\frac{\partial \phi}{\partial y}\right| \le \sqrt{2 osc_y \phi + \delta_2^2}. \tag{5.2}$$

Proof. We follow the analysis of Lemma 3.1, with the main difference being in the construction of the barrier h. Specifically we set $v(x, y, \xi) \equiv \phi(x, y) - \phi(\xi, y)$ and define $h(\tau)$ as the solution of the problem

$$h''(\tau) = -1, \quad h(0) = 0, \quad h(\tau^*) = osc_x \phi,$$

where τ^* will be defined later. We need to compare the functions v and $h(x - \xi)$ in the prism $P_2 \cap \{x - \xi < \mathcal{L}\}$, where

$$P_2 = \left\{(x, \xi, y) : |x| < \frac{\mathcal{L}}{2}, |\xi| < \frac{\mathcal{L}}{2}, |y| < \frac{\mathcal{W}}{2}, \tau^* > x - \xi > 0\right\},$$

whose cross-section along the y-axis is given in Figure 5.1. Obviously (see (3.7)) we have $\Delta(v - h) \ge 0$ in $P_2 \cap \{x - \xi < \mathcal{L}\}$, and hence the maximum is not achieved in the interior of $P_2 \cap \{x - \xi < \mathcal{L}\}$. We need to check the boundary. When $x = \xi$ and $y \in [-\frac{\mathcal{W}}{2}, \frac{\mathcal{W}}{2}]$ we have $v - h = 0$.

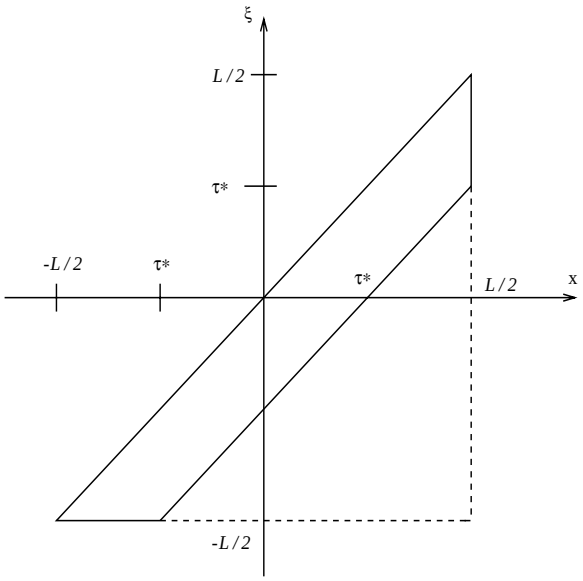


FIG. 5.1. *A cross-section of domain P_2 along a plane in the y-direction.*

For $x - \xi = \tau^*$ we obtain $v - osc_x\phi \leq 0$ for $y \in [-\frac{\mathcal{W}}{2}, \frac{\mathcal{W}}{2}]$. For $x = \frac{\mathcal{L}}{2}$, $\xi \in (-\tau^* + \frac{\mathcal{L}}{2}, \frac{\mathcal{L}}{2})$, and $y \in [-\frac{\mathcal{W}}{2}, \frac{\mathcal{W}}{2}]$ we have that

$$\frac{\partial(v-h)}{\partial x}\bigg|_{x=\frac{\mathcal{L}}{2}} = H + \alpha_2 - h'\left(\frac{\mathcal{L}}{2} - \xi\right).$$

Similarly if $\xi = -\frac{\mathcal{L}}{2}$, $x \in (-\frac{\mathcal{L}}{2}, -\frac{\mathcal{L}}{2} + \tau^*)$, and $y \in [-\frac{\mathcal{W}}{2}, \frac{\mathcal{W}}{2}]$, then

$$\frac{\partial(v-h)}{\partial \xi}\bigg|_{\xi=-\frac{\mathcal{L}}{2}} = -(H - \alpha_1) + h'\left(x + \frac{\mathcal{L}}{2}\right).$$

Therefore if $h' > H + \alpha_2$ (note that $H, \alpha_1, \alpha_2, \delta_1, \delta_2$ are positive constants), then

$$\frac{\partial(v-h)}{\partial x}\bigg|_{x=\frac{\mathcal{L}}{2}} < 0, \qquad \frac{\partial(v-h)}{\partial \xi}\bigg|_{\xi=-\frac{\mathcal{L}}{2}} > 0,$$

and hence we do not have a maximum on these parts of the boundary of P_2. Furthermore, since for $y = \pm\frac{\mathcal{L}}{2}$, $x \in (-\frac{\mathcal{L}}{2}, \frac{\mathcal{L}}{2})$, and for $\xi \in (-\frac{\mathcal{L}}{2}, \frac{\mathcal{L}}{2})$ and $\xi \in (-\frac{\mathcal{L}}{2}, \frac{\mathcal{L}}{2})$ and $0 < x - \xi < \tau^*$ we have $\frac{\partial(v-h)}{\partial y} = 0$, we conclude (see Lemma 3.4 in [8]) that we do not have a maximum here either. It remains to choose τ^* such that $h'(\tau) > H + \alpha_2$ for $\tau \in [0, \tau^*]$. For this we get

$$\tau^* < -(H + \alpha_2) + \sqrt{(H + \alpha_2)^2 + 2osc_x\phi}.$$

As previously we have that

$$|\phi_x(x, y)| \leq h'(0) = \frac{osc_x\phi}{\tau^*} + \frac{\tau^*}{2}.$$

It can be seen that the minimum of $h'(0)$ with respect to τ^* is achieved when $\tau^* = -(H + \alpha_2) + \sqrt{(H + \alpha_2)^2 + 2osc_x\phi}$, from which we obtain relation (5.1).

For $\tau^* \geq \mathcal{L}$ the only difference is the absence of the boundary $x - \xi = \tau^*$, and the boundaries $x = \frac{\mathcal{L}}{2}$, $\xi \in (-\tau^* + \frac{\mathcal{L}}{2}), \frac{\mathcal{L}}{2}$, $y \in [-\frac{\mathcal{W}}{2}, \frac{\mathcal{W}}{2}]$ become $x = \frac{\mathcal{L}}{2}$, $\xi \in (-\frac{\mathcal{L}}{2}, \frac{\mathcal{L}}{2}$, $y \in [-\frac{\mathcal{W}}{2}, \frac{\mathcal{W}}{2}]$, and $\xi = -\frac{\mathcal{L}}{2}$, $x \in (-\frac{\mathcal{L}}{2}, -\frac{\mathcal{L}}{2} + \tau^*)$, $y \in [-\frac{\mathcal{W}}{2}, \frac{\mathcal{W}}{2}]$ become $\xi = -\frac{\mathcal{L}}{2}$, $x \in (-\frac{\mathcal{L}}{2}, \frac{\mathcal{L}}{2})$, $y \in [-\frac{\mathcal{W}}{2}, \frac{\mathcal{W}}{2}]$. We work similarly for the case $\tau^* < \mathcal{L}$

One can obtain relation (5.2) working similarly for the y variable. □

Remark. We conclude the discussion in this section by observing that in the case where $\alpha_1 = \alpha_2 = \alpha$ and $\delta_1 = \delta_2 = \delta$ the numerical experiments that we have conducted, shown in section 7, indicate that a solution of the PDE problem (2.1)–(2.2) is symmetric along the axis $y = 0$. To obtain this solution one can reduce the PDE problem (2.1) to the domain defined by $(-\frac{\mathcal{L}}{2}, \frac{\mathcal{L}}{2}) \times (-\frac{\mathcal{W}}{2}, 0)$ together with the boundary conditions $\frac{\partial \phi(x,y)}{\partial x} = H \pm \alpha$ on $x = \pm\mathcal{L}/2$, $\frac{\partial \phi(x,y)}{\partial y} = \delta$ on $y = -\mathcal{W}/2$, and $\frac{\partial \phi(x,y)}{\partial y} = 0$ on $y = 0$. The solution to the original problem is obtained by applying symmetry across the y-axis. For the reduced problem the estimates (3.9) and (4.8) obtained above can be improved to become

$$-\frac{\mathcal{W}}{2} - \frac{2\delta}{\mathcal{W}} y \leq \frac{\partial \phi(x,y)}{\partial y} \leq \frac{\mathcal{W}}{2} - \frac{2\delta}{\mathcal{W}} y \tag{5.3}$$

and

$$\begin{aligned} &\int_{-\frac{\mathcal{L}}{2}}^{\frac{\mathcal{L}}{2}} \int_{-\frac{\mathcal{W}}{2}}^{0} \left[\left(\frac{\partial^2 \phi}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 \phi}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 \phi}{\partial y^2} \right)^2 \right] dxdy \\ &\leq \frac{1}{2} \mu(\Omega_j) \left[1 + 4 \left(\frac{\alpha}{\mathcal{L}} + \frac{\delta}{\mathcal{W}} \right) \right] + 2\alpha \frac{H\mathcal{W}}{\mathcal{L}} + 5\delta^2 \frac{\mathcal{L}}{\mathcal{W}} + 2\alpha\delta, \end{aligned} \tag{5.4}$$

respectively.

In the particular case when $H = \delta = 0$ and the junction is placed symmetrically inside Ω, we have observed the existence of a solution for which the phase is equal to a constant along the line $x = 0$ (see the top of Figure 7.2). Although we are unable to prove their existence, such solutions have been observed in practice, and their physical justification is well established in the case where no extra fluxons have entered the interior of the window [4, 5]. For this type of solution we are able to obtain the following estimations of its size.

LEMMA 5.2. *For any classical solution $\phi(x, y)$ of the problem (2.1)–(2.2) with $H, \delta = 0$ and $\alpha_i = a$, for which $\phi = k$, k is a constant, at $x = 0$, we have that*

$$k + \frac{x^2}{2} + \left(\frac{\mathcal{L}}{2} - \alpha \right) x \leq \phi \leq k - \frac{x^2}{2} - \left(\alpha + \frac{\mathcal{L}}{2} \right) x \quad \text{for} \quad -\frac{\mathcal{L}}{2} \leq x \leq 0, \tag{5.5}$$

$$k + \frac{x^2}{2} - \left(\frac{\mathcal{L}}{2} - \alpha \right) x \leq \phi \leq k - \frac{x^2}{2} + \left(\frac{\mathcal{L}}{2} + \alpha \right) x \quad \text{for} \quad 0 \leq x \leq \frac{\mathcal{L}}{2}. \tag{5.6}$$

Proof. To obtain relation (5.5) we consider the domain $\Omega_1 \equiv (-\frac{\mathcal{L}}{2}, 0) \times (-\frac{\mathcal{W}}{2}, \frac{\mathcal{W}}{2})$, and from (3.1) we have

$$-\beta - 1 \leq \frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} \leq -\beta + 1,$$

where $\beta = \frac{2\alpha}{\mathcal{L}}$. We define now the function $g(x) \equiv \frac{1-\beta}{2}(x + \frac{\mathcal{L}}{2})^2 + \epsilon x$, where $\epsilon > 0$ and $v \equiv u - g$. Obviously we have that

$$\frac{\partial^2 v(x,y)}{\partial x^2} + \frac{\partial^2 v(x,y)}{\partial y^2} \leq 0;$$

thus v does not achieve its minimum in Ω_1 (unless it is a constant). It does not achieve it on the boundary lines $x = -\frac{\mathcal{L}}{2}$, $y \in [-\frac{\mathcal{W}}{2}, \frac{\mathcal{W}}{2}]$ (since the derivative in the x-direction is negative) and $y = \pm\frac{\mathcal{W}}{2}$ for $x \in (-\frac{\mathcal{L}}{2}, 0)$ (see Lemma 3.4, p. 34 in [9]) either. So, we conclude that the minimum of v occurs at $x = 0$, and we have that

$$v \geq \min\,(u-g)|_{x=0} = k - \frac{\mathcal{L}^2}{8},$$

and therefore

$$u \geq k + \frac{1-\beta}{2}x^2 + (1-\beta)\frac{\mathcal{L}}{2}x - \frac{\beta}{8}\mathcal{L}^2 + \epsilon x.$$

We now set $\hat{v} \equiv u - \hat{g}$, where $\hat{g}(x) \equiv -\frac{\beta+1}{2}(x + \frac{\mathcal{L}}{2})^2 - \epsilon x$, where $\epsilon > 0$, from which we have that

$$\frac{\partial^2 \hat{v}(x,y)}{\partial x^2} + \frac{\partial^2 \hat{v}(x,y)}{\partial y^2} \geq 0.$$

Using arguments similar to those above we can show that $\hat{v} \leq \max \hat{v}|_{x=0}$ to obtain

$$u \leq k - \frac{\beta+1}{2}x^2 - (\beta+1)\frac{\mathcal{L}}{2}x - \frac{\beta}{8}\mathcal{L}^2 - \epsilon x.$$

To conclude the proof of relation (5.5) we simply repeat the above analysis for the domain $(0, \frac{\mathcal{L}}{2}) \times (-\frac{\mathcal{W}}{2}, \frac{\mathcal{W}}{2})$, use the fact that ϵ is an arbitrary positive constant, and simply go from the function u to the function ϕ. □

6. Linearization. For the numerical solution of the semilinear elliptic PDE problem (2.1)–(2.2) one can linearize the PDE equation by means of the following fixed point iteration scheme:

$$L\phi^{(i)} \equiv \Delta\phi^{(i)} - \mathcal{I}_j r\phi^{(i)} = \mathcal{I}_j\left(\sin(\phi^{(i-1)}) - r\phi^{(i-1)}\right), \quad i = 1, 2, \ldots, \tag{6.1}$$

where $r \equiv r(x,y)$ is a relaxation function to accelerate the convergence, and it can be any nonzero function. We start these iterations using an initial guess $u^{(0)}$ of the solution u obtained using one of the approaches described in [4], and we terminate them when the `max`-norm of the difference of two successive approximations of the solution vector ($||\phi^{(i)} - \phi^{(i-1)}||_\infty$) or the `max`-norm of the residual of the problem ($||\Delta\phi^{(i)} - \mathcal{I}\sin\phi^{(i)}||_\infty$) is less than a given tolerance. Two obvious choices for that parameter are $r(x,y) = c$ (constant function) and $r(x,y) = \cos(\phi^{(i-1)}(x,y))$. With the latter one, the iteration scheme (6.1) reduces to the well-known Newton iterative method [14]. The implementation and the performance of this quadratically converging method is given in [4], and its convergence analysis is under way and will be presented elsewhere. For the convergence of (6.1) when r is a positive constant we have the following theorem.

THEOREM 6.1. *If $c \equiv c(r)$ is the measure of the smallest eigenvalue of the operator L, then the iterative method* (6.1) *converges, from any initial guess $\phi^{(0)}$, to the solution of* (3.1)–(3.2) *if*

$$\frac{1}{c}\left(\frac{1}{2} + r\right) < 1. \tag{6.2}$$

Proof. If we denote by $e^{(i)} \equiv \phi - \phi^{(i)}$ the error at the ith iteration, we see that for $i = 1, 2, \ldots$ we have

$$Le^{(i)} = \mathcal{I}_j \left[\cos\left(\frac{e^{(i-1)}}{2} + \phi \right) \sin\left(\frac{e^{(i-1)}}{2} \right) - re^{(i-1)} \right],$$

from which we obtain

$$||Le^{(i)}|| \leq \left\| \cos\left(\frac{e^{(i-1)}}{2} + \phi \right) \sin\left(\frac{e^{(i-1)}}{2} \right) \right\| + r||e^{(i-1)}||.$$

By expanding the sine term and dropping the cosine term in the above relation we have for $i = 1, 2, \ldots$ that

$$||Le^{(i)}|| \leq \left(\frac{1}{2} + r \right) ||e^{(i-1)}||,$$

from which relation (6.2) can be easily obtained by requiring the amplification factor to be less than 1. □

It is worth pointing out here that the lack of convergence, as one increases the current, of the Newton iterative scheme defined above reflects the dynamical instability of the static solution in the time-dependent sine-Gordon system.

7. Numerical experiments and physical relevance. Using the proposed PDE model we have built a powerful simulation tool that accurately and effectively models window Josephson junctions. Our implementation, described in detail in [4], is based on the ELLPACK infrastructure [16], and its basic components are as follows: a uniform discretization of the domain Ω using a tensor product of $n \times n$ grid lines, the Newton linearization scheme, and the discretization of the PDE problem (3.1)–(3.2) using the standard 5-point-star finite difference method. For all experiments we have used a junction $\Omega = [0, 12] \times [0, 3]$, and unless otherwise stated the window Ω_j has sizes $\ell = 10$ and $w = 1$ in the x- and y-direction, respectively, and is placed in the center of Ω. The boundary conditions were selected such that $\alpha_1 = \alpha_2 \equiv \alpha$ and $\delta_1 = \delta_2 \equiv \delta$.

In Figure 7.1 we present the properties of the Newton linearization scheme with which we obtained all the numerical data reported here. On the left we see the history of convergence during the first four iterations. Specifically we plot, in log–log scale, the quantity $||\phi^{(i)} - \phi^{(i-1)}||_\infty$ versus the iteration number i for $i = 1, 2, 3, 4$ with $n = 20, 40$, and 60, and we easily see the quadratic rate of convergence. To measure the accuracy obtained in the fourth iteration, we plot in the middle panel the infinity and the L_2 norms of the residual $(\Delta\phi - \mathcal{I}_j \sin\phi)$ versus n in semilog scale. The theoretically expected [16] second order convergence, with respect to discretization stepsize, of the 5-point-star discretization scheme used to solve the linear problems at every step in (6.1) can be easily verified. The time complexity of the Newton iterative algorithm is presented in the right panel, where we plot the per-iteration CPU time required versus n. As is easily seen this is approximately n^3.

To understand the structure of the solutions for various boundary conditions and confirm the obtained barrier functions, we give in Figures 7.2 and 7.3 a series of contour and three-dimensional plots of the computed solutions and their gradients for three different boundary conditions. Figure 7.2 corresponds to a situation where $H = 0$, $\delta \equiv 0$, and $\alpha \equiv 0$ in the top plate and bottom plate, respectively. In this case the solution has an oscillation in x smaller than 2π.

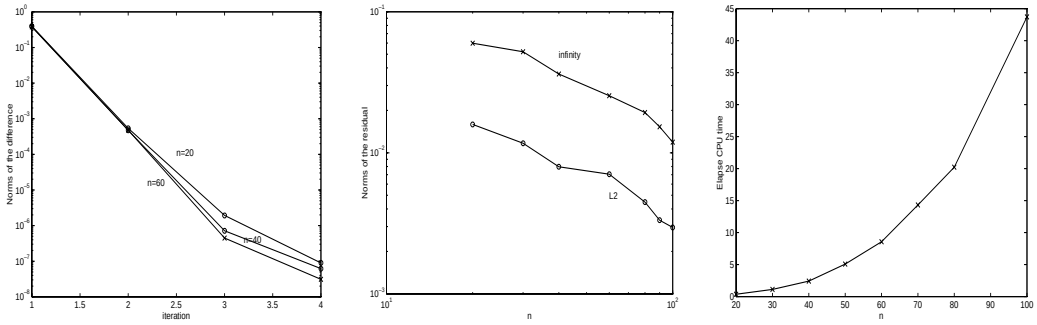


FIG. 7.1. *Log–log plot of the infinity norm of the difference of two successive iterants of Newton method for discretization parameter $n = 20, 40, 60$ versus the iteration number (left), semilog plot of the infinity and L_2 norms of the residual versus the discretization parameter n (middle), and plot of the per-iteration CPU time versus the discretization parameter n (right).*

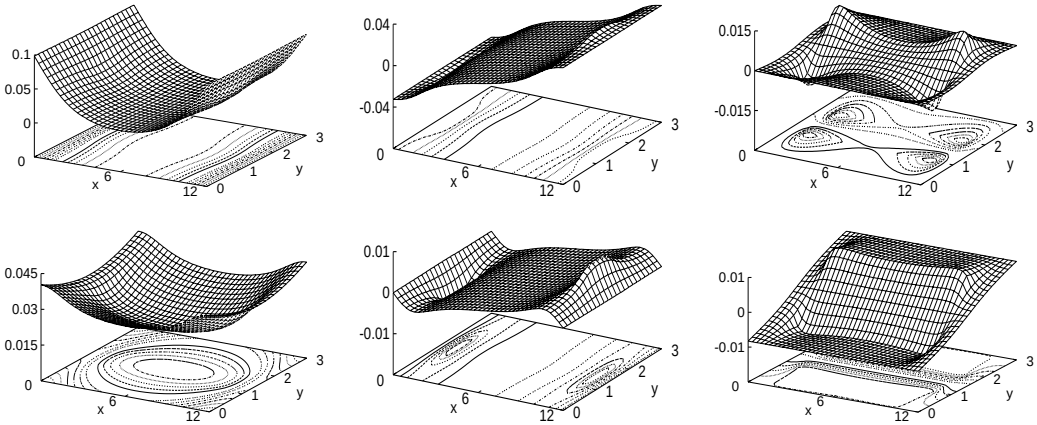


FIG. 7.2. *Plots of the solution (left column) and the derivatives in the x- (middle column) and the y- (right column) directions for $H = 0$ and $\alpha = .033$, $\delta = 0$ (top) and $\alpha = 0$, $\delta = .0083$ (bottom).*

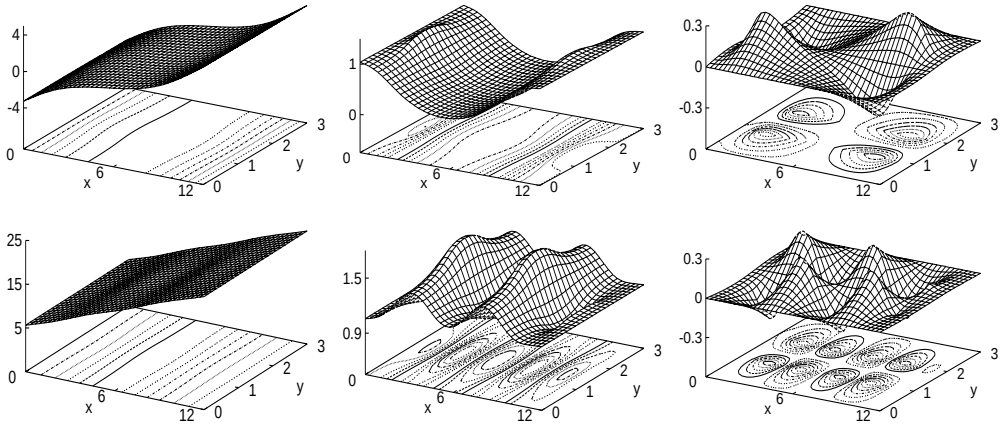


FIG. 7.3. *Plots of the solution (left column) and the derivatives in the x- (middle column) and the y- (right column) directions for $H = 1.1$ $\alpha = .05$ and $\delta = 0$ for one-fluxon (top) and three-fluxon (bottom) solutions.*

TABLE 7.1
A priori bounds for the gradients of the solution associated with the PDE problems considered in Figures 7.2 *and* 7.3.

Figures	Lemma 3.1	Lemma 5.1
7.2 top	$\frac{\partial\phi}{\partial x} \le 12 + 0.055(x-6)$ $\frac{\partial\phi}{\partial y} \le 3$	$\frac{\partial\phi}{\partial x} \le 0.448$ $\frac{\partial\phi}{\partial y} \le 0.14$
7.2 bottom	$\frac{\partial\phi}{\partial x} \le 12$ $\frac{\partial\phi}{\partial y} \le 3 + 0.00553y$	$\frac{\partial\phi}{\partial x} \le 0.223$ $\frac{\partial\phi}{\partial y} \le 0.173$
7.3 top	$\frac{\partial\phi}{\partial x} \le 13.1 + 0.00833(x-6)$ $\frac{\partial\phi}{\partial y} \le 3$	$\frac{\partial\phi}{\partial x} \le 3.72$ $\frac{\partial\phi}{\partial y} \le 0.447$
7.3 bottom	$\frac{\partial\phi}{\partial x} \le 13.1 + 0.055(x-6)$ $\frac{\partial\phi}{\partial y} \le 3$	$\frac{\partial\phi}{\partial x} \le 6.43$ $\frac{\partial\phi}{\partial y} \le 0.447$

For a larger value of the magnetic field $H = 1.1$, shown in Figure 7.3, the oscillation in x of the solution increases. The existence of more than one solution for certain values of the boundary conditions is confirmed here. The first solution presented in the top plate has an oscillation of less than 2π (one-fluxon solution) while the oscillation of the second one is between 4π and 6π (three-fluxon solution).

For the PDE problems considered in Figures 7.2 and 7.3 we present in Table 7.1 the a priori estimates for the gradients of the solution theoretically obtained using Lemmas 3.1 and 5.1, respectively. The confirmation of these lemmas can be readily obtained by comparing the entries of the table with the associated plots in the figures. We also easily see the improvement of the estimates in the x-direction obtained in section 5. To confirm the theoretically obtained estimates in Lemma 5.2 of the solution (in the special case where it has a constant value on the line $x = 0$) we have computed using this lemma the upper and lower bounds of ϕ for the problems considered in the top of Figure 7.2,

$$5.967(x-6) + 0.5(x-6)^2 \le \phi \le -6.033(x-6) - 0.5(x-6)^2 \quad \text{for } 0 \le x \le 6$$

and

$$-5.967(x-6) + 0.5(x-6)^2 \le \phi \le 6.033(x-6) - 0.5(x-6)^2 \quad \text{for } 6 \le x \le 12,$$

and in the top of Figure 7.3,

$$5.95(x-6) + 0.5(x-6)^2 \le \phi \le -6.05(x-6) - 0.5(x-6)^2 \quad \text{for } 0 \le x \le 6$$

and

$$-5.95(x-6) + 0.5(x-6)^2 \le \phi \le 6.05(x-6) - 0.5(x-6)^2 \quad \text{for } 6 \le x \le 12.$$

As is easily seen these estimates agree with the numerical data presented in the associated figures.

As mentioned in section 2 an important question from both a theoretical and a practical point of view is, For what values of H, α, and δ does the solution to our PDE problem exist? Or, equivalently, Which is the maximum current $I_t = 2(\alpha\mathcal{W} + \delta\mathcal{L})$ that the device can carry for a given magnetic field? We have numerically determined

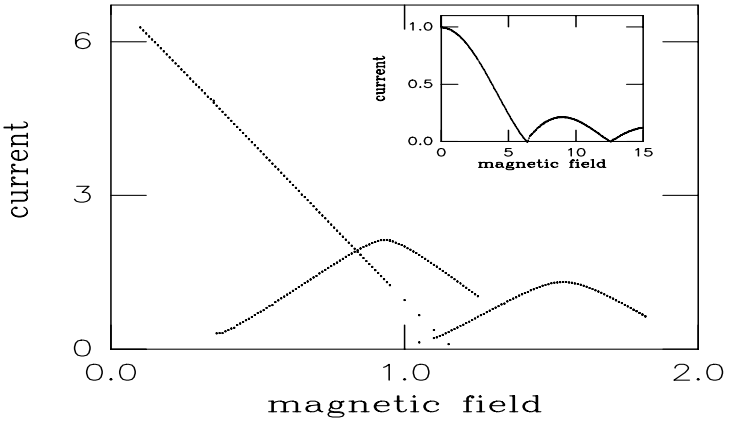


FIG. 7.4. *Allowed values for magnetic field and current for the inline geometry ($\delta \equiv 0$).*

TABLE 7.2
Oscillations for the solutions corresponding to the branches of Figure 7.4 *for several values of the magnetic field H.*

H	I_t	osc_x	osc_y
10^{-2}	6.65	3.73	0.21
0.41	4.37	4.28	0.21
0.47	0.52	7.75	0.21
0.81	2.13	5.51	0.21
1.01	1.97	11.54	0.20
1.21	1.23	12.89	0.19

the relation between the magnetic field and the maximum current for the case where $\delta = 0$, and we present it graphically in Figure 7.4. It is important to note that for pairs of currents and magnetic fields below each (starting from the leftmost) of the three "maximum lines" shown, there exist one-fluxon, two-fluxon, and three-fluxon solutions, respectively. Above them no solutions exist. The overlap of the branches corresponding to one fluxon and three fluxons is consistent with the observation made from Figure 7.3 on the coexistence of a one-fluxon solution and three-fluxon solution for $H = 1.1$. Notice also that in this case the maximum current which is obtained for $H = 0$ is significantly lower than the bound given by (2.3), which is $l \times w = 10$.

We have calculated the oscillations in the x- and y-directions for the solutions corresponding to the maximum current for several values of the magnetic field and reported them in Table 7.2. An initial observation is that the oscillations in the y-direction are small and do not vary significantly as a function of H for the values considered. This indicates that a one-dimensional description of this problem could be possible; such a heuristic approach based on an appropriate rescaling of a one-dimensional sine-Gordon equation is currently under way. In turn the oscillations in the x-direction vary from π to 6π and correspond to the different fluxon branches described in the introduction. Notice, however, that the oscillation for the solution at the right-hand tip of the first branch is larger than 2π, contrary to what happens for the pure one-dimensional sine-Gordon equation. This is due to the fact that the

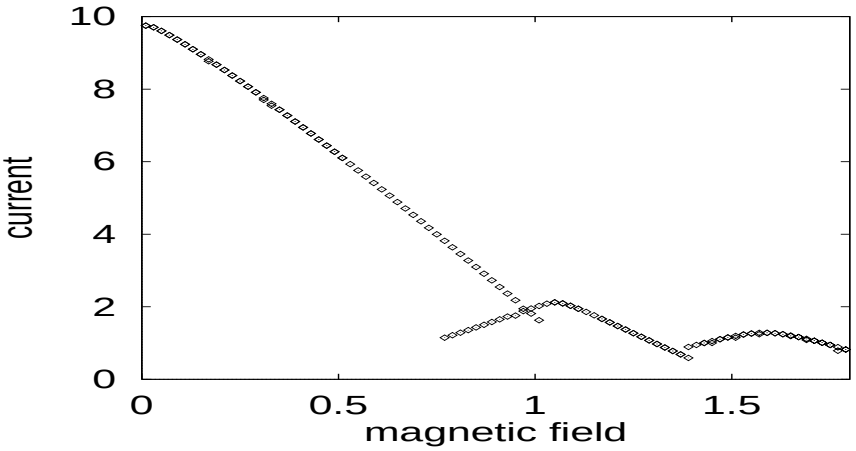


FIG. 7.5. *Allowed values for magnetic field and current for the overlap geometry ($\alpha \equiv 0$).*

TABLE 7.3
Oscillations for the solutions corresponding to the branches of Figure 7.5 *for several values of the magnetic field H.*

H	I_t	osc_x	osc_y
10^{-2}	9.75	0.91	0.54
0.41	6.94	3.05	0.45
0.81	3.45	5.43	0.33
0.87	1.50	9.43	0.24
1.11	1.94	11.87	0.27
1.61	1.27	18.75	0.21

junction domain is smaller than Ω. The case of a very small domain $\Omega = [0,3] \times [0,3]$ and a window $\Omega_j = [1,2] \times [1,2]$ is presented in the inset of Figure 7.4. Another interesting feature is that the branches do not overlap and that their graph is very well approximated by $\left|\frac{\sin \frac{H}{2}}{\frac{H}{2}}\right|$, a feature that is well known for small Josephson junctions [2]. In this situation, the maximum current for $H = 0$ is $l \times w = 1$, corresponding to a solution equal to $\frac{\pi}{2}$ inside the junction.

Returning to the long junction, we have calculated the maximum current when the distribution is of the overlap type ($\alpha = 0$) and present it in Figure 7.5. In this case the maximum current for $H = 0$ is very close to the theoretical bound $l \times w = 10$, indicating that the solution inside the junction is very close to $\frac{\pi}{2}$. As in the inline case the branches overlap, indicating a multiplicity of solutions. The oscillations are reported in Table 7.3. Contrary to the inline case discussed above the oscillation in the y-direction varies significantly, indicating a stronger two-dimensional variation of the solution. As expected the values of the magnetic field corresponding to the zeros of the current coincide in Figures 7.4 and 7.5.

It is possible to use the bounds obtained on $\frac{\partial \phi}{\partial x}$ and $\frac{\partial \phi}{\partial y}$ in section 5 to obtain an estimate of the total current I_t. To do this notice by integrating the PDE (2.1)–(2.2)

over the domains Ω_j and Ω that

$$I_t = 2(\alpha\mathcal{W} + \delta\mathcal{L}) = \int_{\Omega_j} \sin\phi dxdy = \int_{\partial\Omega_j} \nabla\phi \mathbf{n} ds = \int_{\partial\Omega_j} \left(\frac{\partial\phi}{\partial x}\mathrm{n_x} + \frac{\partial\phi}{\partial y}\mathrm{n_y}\right) ds, \tag{7.1}$$

where the last integral is a flux integral taken on the boundary of the junction $\partial\Omega_j$, and $\mathbf{n} = \binom{\mathrm{n_x}}{\mathrm{n_y}}$ is the normal vector associated to this boundary. This integral can be bounded in the case of a junction with arbitrary shape and perimeter P,

$$\left| \int_{\partial\Omega_j} \left(\frac{\partial\phi}{\partial x}\mathrm{n_x} + \frac{\partial\phi}{\partial y}\mathrm{n_y}\right) ds \right| \leq \left(\max\left|\frac{\partial\phi}{\partial x}\right| + \max\left|\frac{\partial\phi}{\partial y}\right| \right) P,$$

so that using Lemma 5.2 one obtains the following inequality for α and δ assumed positive:

$$\frac{I_t}{2} = \alpha\mathcal{W} + \delta\mathcal{L} \leq \frac{P}{2}\left(\sqrt{2osc_x + (H+\alpha)^2} + \sqrt{2osc_y + \delta^2}\right). \tag{7.2}$$

In the case of a rectangular junction centered in the domain Ω, this estimate can be improved by separating the integrals on the parts of the boundary $\partial\Omega_j$ parallel to the x- and y-directions to obtain

$$\begin{aligned}\int_{\partial\Omega_j} \left(\frac{\partial\phi}{\partial x}\mathrm{n_x} + \frac{\partial\phi}{\partial y}\mathrm{n_y}\right) ds = &\int_{-\frac{l}{2}}^{\frac{l}{2}} \left[\frac{\partial\phi}{\partial y}\left(x, \frac{w}{2}\right) - \frac{\partial\phi}{\partial y}\left(x, -\frac{w}{2}\right)\right] dx \\ &+ \int_{-\frac{w}{2}}^{\frac{w}{2}} \left[\frac{\partial\phi}{\partial x}\left(\frac{l}{2}, y\right) - \frac{\partial\phi}{\partial x}\left(-\frac{l}{2}, y\right)\right] dy.\end{aligned}$$

Recall that l and w are the length and width of the junction domain Ω_j, respectively. One can then bound the absolute values of the above integrals using Lemma 5.2 and obtain

$$\frac{I_t}{2} = \alpha\mathcal{W} + \delta\mathcal{L} \leq \sqrt{2osc_x + (H+\alpha)^2}\, w + \sqrt{2osc_y + \delta^2}\, l. \tag{7.3}$$

This upper bound for the current is not as sharp as (2.3). For example, for the case of Figure 7.4 for $H = 0.41$ we find $I_t = 4.37$ corresponding to $\alpha = 0.728$. Using the values of the oscillations given by Table 7.2 we obtain for the right-hand side of the inequality (7.3) 9.62, which corresponds to a total current of 19.24, while the maximum current allowed is $l \times w = 10$.

8. Conclusions and future work. Josephson junctions have already proved themselves to be technologically useful, and it is our belief that their importance will increase significantly in the near future. Many recent reports and books have been dedicated to the analysis of the one-dimensional case, where one can usually give the solution of the associated boundary value problem analytically in terms of elliptic functions. Our report is, to the best of our knowledge, the first to try to theoretically analyze the semilinear PDE problem that effectively and accurately models two-dimensional window Josephson junctions. Specifically we established the existence of solutions and obtained regularity and a priori estimates for the derivatives of the solution. We had to use a specific method to establish these estimates instead of the well-known theory for elliptic PDEs [9] because the solution is defined modulo a

multiple of 2π due to the periodicity of the nonlinearity and the Neumann boundary conditions, and therefore its norm cannot be bounded.

From the practical point of view this study validates the practical observation that for a pure junction ($\Omega_j \equiv \Omega$) of small dimensions $\mathcal{L} < 2$, $\mathcal{W} < 2$ and zero boundary conditions the only solutions are the constants $n\pi$, where n is an integer. Another important practical result is that if I_j depends only on x and $\delta \equiv 0$ and if $\mathcal{W} < 2$, then $\frac{\partial \phi}{\partial y} \equiv 0$. It is interesting to notice that in both results, the value 2 comes up. The same value appears in the one-dimensional reductions of the problem, where it corresponds to the maximum of $\frac{\partial \phi}{\partial x}$ for the separatrix of the pendulum phase space.

The theoretical analysis of two-dimensional window Josephson junctions is by no means complete. Below are some of the issues that are of practical interest (and as such some experimental analysis has already been carried out), and their theoretical analysis will be mathematically challenging.

Notice that all a priori estimates obtained are independent of the window Ω_j. This is due to the fact that we bound $|\sin(\phi)|$ by 1 very early in our analysis. Therefore, although these estimates seem to be very generous for the PDE problems considered in Table 7.1, they are sharp for large windows. Nevertheless, since many important physical properties of Josephson junctions depend on the size and the geometry of the window [4, 5] new a priori estimates which sense the geometrical parameters of the window would be of importance.

The maximum current that a Josephson junction can carry for a given configuration of α's and δ's and for a given H is another point of interest, and its theoretical estimation is a challenging and difficult problem. One approach for that is carrying out a three-parameter stability analysis. These parameters are the values at the boundary conditions and the size of the window. Such stability analysis to determine the turning and bifurcation points and eigenvalues corresponding to the different solutions is under way.

The method of Newton proved to be a very reliable and efficient linearization tool. We believe that the proof of its quadratic convergence at continuum level (PDE analysis) or discrete level (numerical analysis) is another interesting mathematical problem. This problem does not have a unique solution and is therefore ill-posed in the Hadamard sense.

Appendix. Proof of existence of the generalized solution for the auxiliary problem. To apply the Leray–Schauder theorem we consider the Banach space $\mathcal{B}$,

$$\mathcal{B} = \left\{ u \in C^1(\bar{\Omega}) \ \text{ and } \ \int_\Omega u dx dy = 0 \right\},$$

and construct the mapping

$$\kappa T : \forall u \in \mathcal{B} \longrightarrow w,$$

where w is the solution of the problem

$$\Delta w = \kappa \left[\mathcal{I}_j \sin(u + \zeta + f) - \frac{1}{\mu(\Omega)} \int_{\Omega_j} \sin(u + \zeta + f) dx dy \right] \equiv \mathcal{F}(x, y) \quad \text{in } \Omega,$$

$$\frac{\partial w}{\partial n} = 0 \quad \text{on } \partial\Omega.$$

Note that we have formed the right-hand side $\mathcal{F}$ so that the solution of the above linear PDE problem exists up to a constant and, as it can be easily seen using the a priori estimates obtained in section 3, it belongs to $C^{1,\gamma}(\bar{\Omega})$. From this class of infinitely many solutions, we can select (by choosing the appropriate constant) the one that satisfies relation (4.7). Therefore we have constructed a mapping from $\mathcal{B}$ to $\mathcal{B}^\gamma$, where $\mathcal{B}^\gamma = \left\{u \in C^{1,\gamma}(\bar{\Omega}) \text{ and } \int_\Omega u dx dy = 0\right\}$. The mapping $T : \mathcal{B} \longrightarrow \mathcal{B}^\gamma$ is bounded and hence $T : \mathcal{B} \longrightarrow \mathcal{B}$ is compact. To apply the Leray–Schauder theorem, and therefore to prove the existence of a fixed point of T, we need only show that for every solution of $w = \kappa T w$, $\kappa \in [0, 1]$ we have that $||w||_{C^1(\overline{\Omega})}$ is bounded. This is a direct consequence of the a priori estimates we have already obtained. Note that from Schauder estimates we also have that the above-mentioned solution belongs to $C^3(\Omega)$.

REFERENCES

[1] A. Barone, F. Esposito, K. Likharev, V. Semenov, B. Todorov, and R. Vaglio, *Effect of boundary conditions upon the phase in two-dimensional Josephson junctions*, J. Appl. Phys., 53 (1982), pp. 5802–5809.

[2] A. Barone and G. Paterno, *Physics and Applications of the Josephson Effect*, John Wiley, New York, 1982.

[3] J. Caputo, N. Flytzanis, Y. Gaididei, and E. Vavalis, *Two-dimensional effects in Josephson junctions: Static properties*, Phys. Rev., (1996), pp. 2092–2021.

[4] J. Caputo, N. Flytzanis, and E. Vavalis, *A semi-linear elliptic PDE model for static solution of Josephson junctions*, Internat. J. Modern Phys. C, 6 (1995), pp. 241–262.

[5] J. Caputo, N. Flytzanis, and E. Vavalis, *Effect of geometry on fluxon width in a Josephson junction*, Internat. J. Modern Phys. C, 7 (1996), pp. 191–216.

[6] J. G. Caputo, N. Flytzanis, and M. Devoret, *Dressed fluxon in a window Josephson junction*, Phys. Rev. B, 50 (1994), pp. 6471–6474.

[7] T. Chow, *Enhancement of the critical current in a Josephson tunneling junction with defect size in micrometers*, Phys. C, 255 (1995), pp. 311–318.

[8] R. Flesch, M. Forest, and A. Sinha, *Numerical inverse spectral transform for periodic sine-Gordon equation: Theta function solutions and their linearized stability*, Phys. D, 48 (1991), pp. 169–231.

[9] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1983.

[10] E. P. Houwman, J. G. Gijsbersten, J. Flokstra, and H. Rogalla, *On the suppression of the sidelobes of the supercurrent in small Josephson*, Phys. C, 164 (1991), pp. 339–344.

[11] S. Kruzhkov, *Quasilinear parabolic equations and systems with two independent variables*, Trudy Sem. Petrovsk Vyp., 5 (1979), pp. 217–272.

[12] O. A. Ladyzhenskaya, *The Boundary Value Problems of Mathematical Physics*, Prentice–Hall, Englewood Cliffs, NJ, 1986.

[13] O. A. Ladyzhenskaya and N. N. Uraltseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[14] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[15] C. Owen and D. Scalapino, *Vortex structure and critical currents in Josephson junctions*, Phys. Rev., 164 (1967), pp. 538–544.

[16] J. Rice and R. Boisvert, *Solving Elliptic Problems Using ELLPACK*, Springer-Verlag, New York, 1985.

[17] A. Tersenov, *On the first boundary value problem for quasilinear parabolic equations with two independent variables*, Arch. Ration. Mech. Anal., 152 (2000), pp. 81–92.

[18] A. Tersenov, *On quasilinear non-uniformly parabolic equations in general form*, J. Differential Equations, 142 (1998), pp. 263–276.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1380–1398

YOUNG MEASURE SOLUTIONS FOR NONCONVEX ELASTODYNAMICS*

MARC OLIVER RIEGER†

Abstract. We study the nonlinear equation of elastodynamics where the free energy functional is allowed to be nonconvex. We define the notion of Young measure solutions for this problem and prove an existence theorem in this class. This can be used as a model for the evolution of microstructures in crystals. We furthermore introduce an optional coupling with a parabolic equation and prove the existence of a Young measure solution for this system.

Key words. nonlinear elasticity, hyperbolic-parabolic systems, Young measures, nonconvex variational problems

AMS subject classifications. 74B20, 74N15, 74N25

PII. S0036141001392141

1. Introduction.

A crucial assumption to obtaining the existence of weak solutions for nonlinear elasticity equations in the static case is the quasi convexity of the underlying free energy potential (see [2]). However, in many cases the quasi convexity of the potential is not appropriate to reflect the physical situation. Therefore a weaker concept for solutions has been introduced, the so-called Young measure solutions (YM-solutions). This concept can be applied to crystals where nonconvex elasticity equations can be used to describe the development of microstructures (which are important especially for shape-memory alloys), as has been pointed out in the fundamental paper [3]. (For further information and references consider, e.g., [16], [15].)

The equation of elastodynamics,

$$u_{tt}(x,t) - \operatorname{div} S(\nabla u(x,t)) = 0,$$

is even more difficult to handle. Global existence results for weak solutions have been found only in one space dimension in [10], [19]. Under certain convexity assumptions Dafermos and Hrusa [5] proved the local existence of smooth solutions.

The concept of YM-solutions has been applied to dynamic problems in [20], [12], and [6] (in the context of the forward-backward heat equation) and was applied to the wave equation by [14] and [7].

An approach to the dynamic elasticity equation (with some additional assumptions on the free energy, valid in particular for antiplane shear, and with an *optional* coupling to a parabolic partial differential equation) was presented in [17] using the method of discretization in time. (For a numerical implementation of this construction, see [18] and [4].) A similar result was obtained in [9], where (in a different context) existence was proved in arbitrary space dimensions for the *polyconvex* case under some growth conditions.

In the first part of this article we prove the existence (globally in time for large initial data) of YM-solutions for nonconvex elasticity equations in arbitrary space

*Received by the editors July 11, 2001; accepted for publication (in revised form) October 4, 2002; published electronically May 12, 2003. This work was partially supported by the Center for Nonlinear Analysis under NSF grant DMS-9803791.

http://www.siam.org/journals/sima/34-6/39214.html

†Center for Nonlinear Analysis, Carnegie Mellon University, Pittsburgh, PA (rieger@andrews.cmu.edu). Formerly at the Max-Planck-Institute for Mathematics in the Sciences, Leipzig, Germany.

dimensions under some growth conditions on the free energy. In contrast to [9] we have to assume that the Andrews–Ball condition (see below) is satisfied, but we do *not* need polyconvexity.

In the second part we study a model problem where we couple a nonconvex elasticity equation with a parabolic equation (possibly of forward-backward type). The physical motivation is to study crystals consisting of different types of atoms, where solid state diffusion occurs and influences the elastic properties of the material. The mathematical structure is also similar to thermoelastic problems (cf. [24]). We extend the concept of YM-solutions to this hyperbolic-parabolic system and prove existence.

We remark that the coupling is not crucial for the proof, so the system studied in section 2 is a special case of the system in section 3, but we need stronger bounds for the energy density in the coupled case. This is probably only due to technical reasons, but we decided to discuss both cases separately.

2. YM-solutions for an elasticity equation. In this section we prove the existence of YM-solutions for nonconvex elasticity equations. Let $p \geq 2$ be a fixed constant. (Later p will denote the growth rate of the free energy at infinity.) By p' we denote its conjugate, i.e., $\frac{1}{p} + \frac{1}{p'} = 1$.

Throughout this article we denote by M a positive generic constant depending only on the initial data. By $||\cdot||$ we denote the $L^2(\Omega)$-norm.

For an open bounded set $\Omega \subset \mathbb{R}^n$ with Lipschitz boundary, $T > 0$, $g \in W^{1,p}(\Omega, \mathbb{R}^m)$ and a function $u : \Omega \to \mathbb{R}^m$ we study the initial boundary value problem

$$\begin{aligned} u_{tt}(x,t) - \operatorname{div} S(\nabla u(x,t)) &= 0, \quad (x,t) \in \Omega \times [0,T), \\ u(\cdot,0) &= u_0, \\ u_t(\cdot,0) &= z_0, \\ u &= g \quad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

with $S = \nabla\phi$ and $\phi \in \mathcal{C}^2(\mathbb{R}^{m\times n}, \mathbb{R}_+)$ satisfying the growth conditions (for positive constants M_1, M_2)

$$\begin{aligned} |S(A)| &\leq M_2(|A|^{p-1} + 1), \\ M_1(|A|^p - 1) \leq \phi(A) &\leq M_2(|A|^p + 1) \end{aligned} \tag{2.2}$$

and S satisfying the Andrews–Ball condition (introduced in [1] and generalized in [11]) for some $R > 0$:

$$(S(F_1) - S(F_2))\,(F_1 - F_2) \geq 0 \tag{2.3}$$

for all $F_1 \in \mathbb{R}^{m\times n}$, $F_2 \in \mathbb{R}^{m\times n}$, and $|F_1|, |F_2| \geq R$. An interpretation of this condition is that for "large" values the potential ϕ is assumed to be convex. The condition is not very restrictive since every sufficiently smooth function ϕ on an arbitrarily large ball $\mathcal{B}(0,R)$ can be extended to a function $\tilde{\phi}$ such that $\tilde{S} := \nabla\tilde{\phi}$ satisfies the Andrews–Ball condition.

We can even relax this condition slightly (see [11]): It is sufficient to assume that there exists a constant $M > 0$ such that for all F_1, $F_2 \in \mathbb{R}^{m\times n}$,

$$(S(F_1) - S(F_2))\,(F_1 - F_2) \geq -M|F_1 - F_2|^2. \tag{2.4}$$

We now want to define a YM-solution. Therefore we introduce a measure ν expressing the probability distribution of the deformation gradient at a certain point $(x,t) \in$

$\Omega \times (0,T)$. For "classical" solutions this measure will be a Dirac measure concentrated in ∇u.

DEFINITION 2.1 (YM-solutions for elasticity). *A pair* (u,ν) *is a* YM-solution *of the system* (2.1) *if for fixed* $T>0$,

$$\begin{aligned} u \in W^{1,\infty}((0,T),L^2(\Omega)), \quad u-g \in L^\infty((0,T),W_0^{1,p}(\Omega)),\\ \nu=(\nu_{x,t})_{x,t} \text{ is a probability measure,}\\ \int_0^T \int_\Omega \langle \nu, S(\cdot)\rangle \nabla\zeta - u_t\zeta_t \,dx\,dt = 0 \quad \forall \zeta \in \mathcal{C}_0^\infty((0,T)\times\Omega),\\ \nabla u(x,t) = \langle \nu_{x,t}, Id\rangle \quad a.e. \end{aligned}$$

Here $\langle \nu, S(\cdot)\rangle$ *is defined as a dual pairing of* S *with the measure* ν, *i.e.,* $\langle \nu, S(\cdot)\rangle := \int S(A)\,d\nu(A)$.

In this section we prove the following existence theorem.

THEOREM 2.1 (existence of YM-solutions). *Assume* $\phi \in \mathcal{C}^2$, *that the growth conditions* (2.2) *are satisfied, and that one of the conditions* (2.3) *or* (2.4) *is valid. Furthermore let* $u_0 - g \in W_0^{1,p}(\Omega)$, $z_0 \in H_0^1(\Omega)$. *Then there exists a YM-solution* (u,ν) *of problem* (2.1).

To prove this we use a viscosity regularization, based on an idea of [21]. Under the assumptions stated above the following viscoelastic equation (together with the standard initial and boundary conditions) has a weak solution (see [11], or consider [8] for more general viscosity terms):

$$u_{tt}^\varepsilon(x,t) - \operatorname{div} S(\nabla u^\varepsilon(x,t)) - \varepsilon \Delta u_t^\varepsilon(x,t) = 0.$$

More precisely there exists

$$\begin{aligned} u^\varepsilon \in W^{2,2}((0,T),W^{-1,p'}(\Omega)) \cap W^{1,2}((0,T),W^{1,2}(\Omega)) \cap W^{1,\infty}((0,T),L^2(\Omega)),\\ u^\varepsilon - g \in L^\infty((0,T),W_0^{1,p}(\Omega)) \end{aligned}$$

such that for all $T>0$ and for all $\zeta \in \mathcal{C}_0^\infty((0,T)\times\Omega))$,

$$\int_0^T \int_\Omega (S(\nabla u^\varepsilon) + \varepsilon \nabla u_t^\varepsilon)\nabla\zeta - u_t^\varepsilon \zeta_t \,dx\,dt = 0. \tag{2.5}$$

Furthermore we have the inequality

$$\frac{1}{2}||u_t^\varepsilon||^2 + ||\nabla u^\varepsilon||_{L^p(\Omega)}^p + \int_0^T ||\sqrt{\varepsilon}\nabla u_t^\varepsilon||^2\,dt \le M,$$

where $M>0$ is independent of ε and t. To get this estimate we can follow [11], where we simply add an ε to the viscosity term. Additionally we use the growth condition on ϕ.

These bounds on u^ε imply that there exists a subsequence, again denoted by u^ε, with

$$u^\varepsilon \overset{\star}{\rightharpoonup} u \quad \text{in } L^\infty((0,T),W^{1,p}(\Omega)) \cap W^{1,\infty}((0,T),L^2(\Omega)),$$

and $(\nabla u^\varepsilon(\cdot,t))_\varepsilon$ generates for every fixed $t\in(0,T)$ a Young measure $\nu_{\cdot,t}$.

Now we claim that (u,ν) is a YM-solution of the elasticity equation. To prove this we consider the convergence of the terms in the viscoelastic equation (taking

subsequences if necessary). First we observe that by the convergence proved above and the Hölder inequality,

$$u_t^\varepsilon \overset{\star}{\rightharpoonup} u_t \quad \text{in} \quad L^2((0,T),L^2(\Omega)).$$

Thus $\int_0^T \int_\Omega u_t^\varepsilon \zeta_t \, dx\, dt$ (the third term in the weak equation (2.5)) converges to $\int_0^T \int_\Omega u_t \zeta_t \, dx\, dt$. On the other hand $\int_0^T \int_\Omega \varepsilon \nabla u_t^\varepsilon \nabla \zeta \, dx\, dt$ converges for $\varepsilon \to 0$ to zero as the following calculation (using the Cauchy–Schwarz inequality) proves:

$$\int_0^T \int_\Omega \varepsilon \nabla u_t^\varepsilon \nabla \zeta \, dx\, dt \le \Bigg(\varepsilon \underbrace{\int_0^T ||\sqrt{\varepsilon} \nabla u_t^\varepsilon||^2 \, dt}_{\le M} \Bigg)^{1/2} \Bigg(\underbrace{\int_0^T ||\nabla \zeta||^2 \, dt}_{=const.} \Bigg)^{1/2} \to 0.$$

It remains to consider the term $\int_0^T \int_\Omega S(\nabla u^\varepsilon) \nabla \zeta \, dx\, dt$. If we define $\nu_{\cdot,t}$ for all $t \in (0,T)$ as the gradient Young measure generated by the sequence $\nabla u^\varepsilon(\cdot,t)$ (for a definition and an existence proof, consider, e.g., [13], [15], or [16]), we can see that $S(\nabla u^\varepsilon(\cdot,t))$ converges for all $t \in (0,T)$ weakly in $L^{p-1}(\Omega)$ to $\langle \nu_{\cdot,t}, S \rangle$.

On the other hand a subsequence of $S(\nabla u^\varepsilon)$ converges weakly-$\star$ in $L^\infty((0,T), L^{p'}(\Omega))$, since the bounds from the energy estimate together with the growth condition imply

$$\begin{aligned} \sup_t ||S(\nabla u^\varepsilon)||^{p'}_{L^{p'}(\Omega)} &\le M \sup_t \int_\Omega (1 + |\nabla u^\varepsilon|^{p-1})^{p'} \, dx \\ &\le M \sup_t \left(1 + \int_\Omega |\nabla u^\varepsilon|^{(p-1)p'} \right) dx \\ &= M \sup_t \left(1 + ||\nabla u^\varepsilon||^p_{L^p(\Omega)} \right) dx \\ &\le M. \end{aligned}$$

Hence the term $S(\nabla u^\varepsilon)$ converges weakly-$\star$ in $L^\infty((0,T), L^{p'}(\Omega))$ to $\langle \nu, S \rangle$, and since $\nabla \zeta \in \mathcal{C}_0^\infty((0,T) \times \Omega) \subset L^1((0,T), L^p(\Omega))$ we have derived that (u, ν) is a YM-solution of the elasticity equation, proving Theorem 2.1. ∎

We notice that the Andrews–Ball condition was used only to find a solution to the viscous system.

3. Hyperbolic-parabolic systems. If we want to consider a coupling between elasticity and diffusion, or if we want to study thermoelastic problems, we have to couple a parabolic equation (possibly of forward-backward type) to the elasticity equation. For this purpose we study the following model problem, where $\Omega \subset \mathbb{R}^n$ is a domain with Lipschitz boundary, $T > 0$, $(x,t) \in \Omega \times [0,T)$, $g \in H^1(\Omega, \mathbb{R}^m)$, $u : \Omega \times [0,T) \to \mathbb{R}^m$, and $c : \Omega \times [0,T) \to \mathbb{R}^d$:

$$\begin{aligned} u_{tt}(x,t) - \operatorname{div} S(\nabla u(x,t), c(x,t)) &= 0, \\ c_t(x,t) - \operatorname{div} K(\nabla c(x,t), u(x,t)) &= 0, \\ u(\cdot,0) &= u_0, \\ u_t(\cdot,0) &= z_0, \\ c(\cdot,0) &= c_0, \\ u &= g \quad \text{on } \partial\Omega, \\ \vec{n} K(\nabla c, u) &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{3.1}$$

with $S = \nabla_1 \phi$ and $K = \nabla_1 \psi$ (∇_1 denoting the derivative with respect to the first variable). By $\vec{n}$ we denote the outward normal on $\partial\Omega$.

Here we consider only the case $p = 2$; i.e., we assume that S and K are of linear growth in the first variable and ϕ, $\psi \in \mathcal{C}^2$ are positive and of quadratic growth in the first variable. More precisely there are constants M_1, $M_2 > 0$ such that for all $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^d$, and $a \in \mathbb{R}^m$ the following estimates hold:

$$\begin{aligned} M_1(|A|^2 - 1) \le \phi(A,b) &\le M_2(|A|^2 + |b|^2 + 1),\\ M_1(|B|^2 - 1) \le \psi(B,a) &\le M_2(|B|^2 + |a|^2 + 1),\\ S(A,b) &\le M_2(|A| + |b| + 1),\\ K(B,a) &\le M_2(|B| + |a| + 1). \end{aligned} \tag{3.2}$$

Furthermore we assume that S and K are globally Lipschitz continuous.

We want to remark that (3.1) is only a model problem for studying some typical mathematical difficulties. A realistic model for diffusion phenomena should include at least a ∇u-dependence of the diffusion tensor K rather than a u-dependence.

We extend the notion of YM-solutions to the coupled system, where the measure ν describes the probability distribution of the gradient of u (in the same way as in the last section) and the measure μ describes the probability distribution of the gradient of c.

Definition 3.1 (YM-solutions for an hyperbolic-parabolic system). *We call the quadruple (u, ν, c, μ) a* YM-solution *of the system* (3.1) *if for $T > 0$,*

$$\begin{aligned} u \in W^{1,\infty}((0,T), L^2(\Omega)), \quad u - g \in L^\infty((0,T), H_0^1(\Omega)),\\ c \in W^{1,2}((0,T), L^2(\Omega)) \cap L^\infty((0,T), H^1(\Omega)),\\ \nu = (\nu_{x,t})_{x,t},\ \mu = (\mu_{x,t})_{x,t},\ \textit{probability measures,} \end{aligned}$$

$$\int_0^T \int_\Omega \langle \nu, S(\cdot, c)\rangle \nabla\zeta - u_t\zeta_t \, dx\, dt = 0 \quad \forall \zeta \in H^1((0,T)\times\Omega),\ \zeta|_{\partial\Omega} = 0, \tag{3.3}$$

$$\int_0^T \int_\Omega \langle \mu, K(\cdot, u)\rangle \nabla\zeta + c_t\zeta \, dx\, dt = 0 \quad \forall \zeta \in H^1((0,T)\times\Omega), \tag{3.4}$$

$$\begin{aligned} \nabla u(x,t) &= \langle \nu_{x,t}, Id\rangle \quad a.e.,\\ \nabla c(x,t) &= \langle \mu_{x,t}, Id\rangle \quad a.e. \end{aligned}$$

In the rest of this section we prove the following existence theorem.

Theorem 3.1 (existence of YM-solutions). *Let S, K, ϕ, ψ satisfy the conditions in* (3.2)*, and assume that S and K are globally Lipschitz continuous. Then for $u_0 - g \in H_0^1$, $z_0 \in H_0^1$, $c_0 \in H^1$, $\vec{n}K(\nabla c_0, 0) = 0$ there exists a YM-solution (u, ν, c, μ) of the problem stated above.*

To prove this theorem we apply the same methods as in the previous section: We first prove the existence of a weak solution for our system equipped with additional dissipation terms; i.e., we study (for $\varepsilon > 0$)

$$\begin{aligned} u^\varepsilon_{tt}(x,t) - \operatorname{div} S(\nabla u^\varepsilon(x,t), c^\varepsilon(x,t)) - \varepsilon\Delta u^\varepsilon_t(x,t) = 0,\\ c^\varepsilon_t(x,t) - \operatorname{div} K(\nabla c^\varepsilon(x,t), u^\varepsilon(x,t)) - \varepsilon\Delta c^\varepsilon_t(x,t) = 0,\\ u^\varepsilon(\cdot,0) = u_0,\\ u^\varepsilon_t(\cdot,0) = z_0,\\ c^\varepsilon(\cdot,0) = c_0, \end{aligned}$$

$$u^\varepsilon = g \quad \text{on } \partial\Omega,$$
$$\vec{n}(K(\nabla c^\varepsilon, u^\varepsilon) + \varepsilon \nabla c_t^\varepsilon) = 0 \quad \text{on } \partial\Omega. \tag{3.5}$$

For this system we can prove the following theorem.

THEOREM 3.2. *For every $T > 0$ and $u_0 - g \in H_0^1$, $z_0 \in H_0^1$, $c_0 \in H^1$, $\vec{n}K(\nabla c_0, 0) = 0$ there exists a weak solution $(u^\varepsilon, c^\varepsilon)$ of the system (3.5); i.e.,*

$$u^\varepsilon \in L^\infty((0,T), H_0^1(\Omega)) \cap W^{1,\infty}((0,T), L^2(\Omega)) \cap W^{1,2}((0,T), H^1(\Omega)) \cap W^{2,2}((0,T), H^{-1}(\Omega)),$$
$$c^\varepsilon \in W^{1,2}(\mathbb{R}^+, L^2(\Omega)) \cap L^\infty(\mathbb{R}^+, H^1(\Omega)),$$

and

$$\int_0^T \int_\Omega S(\nabla u^\varepsilon, c^\varepsilon)\nabla\zeta + \varepsilon \nabla u_t^\varepsilon \nabla\zeta - u_t^\varepsilon \zeta_t \, dx\, dt = 0 \quad \forall \zeta \in H_0^1((0,T)\times\Omega), \tag{3.6}$$

$$\int_0^T \int_\Omega K(\nabla c^\varepsilon, u^\varepsilon)\nabla\zeta + \varepsilon \nabla c_t^\varepsilon \nabla\zeta + c_t^\varepsilon \zeta \, dx\, dt = 0 \quad \forall \zeta \in H^1((0,T)\times\Omega). \tag{3.7}$$

Furthermore we have the following inequality:

$$\frac{1}{2}||u_t^\varepsilon||^2 + ||u^\varepsilon||_{H^1}^2 + ||c^\varepsilon||_{H^1}^2 + \int_0^T ||\sqrt{\varepsilon}\nabla u_t^\varepsilon||^2 \, dt + \int_0^T ||\sqrt{\varepsilon}\nabla c_t^\varepsilon||^2 \, dt \le M. \tag{3.8}$$

For the proof of this theorem we apply the methods introduced by [12] for the heat equation and [7] for the wave equation. These methods were used for viscoelasticity by [8] and [11]. For a different coupled system describing thermoviscoelastic materials, Zimmer proved the existence of weak-renormalized solutions [24].

First we discretize with respect to time. To make life easier we drop the ε in the notation of u^ε and c^ε and use u and c instead. We denote the discretized variables by $(u^{h,j})_{h,j}, (c^{h,j})_{h,j}$. (Often we will drop the h.) For $j = 0, 1, \ldots$ we will construct (weak) solutions $u^{h,j} \in H_0^1(\Omega)$, $c^{h,j} \in H^1(\Omega)$ of these discretized equations (together with the standard boundary conditions), valid in $H^{-1}(\Omega)$:

$$\frac{u^{h,j} - 2u^{h,j-1} + u^{h,j-2}}{h^2} - \operatorname{div} S(\nabla u^{h,j}, c^{h,j-1}) - \varepsilon \frac{\Delta u^{h,j} - \Delta u^{h,j-1}}{h} = 0,$$
$$\frac{c^{h,j} - c^{h,j-1}}{h} - \operatorname{div} K(\nabla c^{h,j}, u^{h,j-1}) - \varepsilon \frac{\Delta c^{h,j} - \Delta c^{h,j-1}}{h} = 0,$$
$$u^{h,0} = u_0, \quad u^{h,-1} = u_0 - hz_0, \quad c^{h,0} = c_0.$$

More precisely we consider the integral form of these equations; i.e., for $\zeta \in H_0^1(\Omega)$, $\xi \in H_0^1(\Omega)$ we have

$$\int_\Omega \frac{u^{h,j} - 2u^{h,j-1} + u^{h,j-2}}{h^2}\zeta + S(\nabla u^{h,j}, c^{h,j-1})\nabla\zeta + \varepsilon \frac{\nabla u^{h,j} - \nabla u^{h,j-1}}{h}\nabla\zeta = 0, \tag{3.9}$$

$$\int_\Omega \frac{c^{h,j} - c^{h,j-1}}{h}\xi + K(\nabla c^{h,j}, u^{h,j-1})\nabla\xi + \varepsilon \frac{\Delta c^{h,j} - \Delta c^{h,j-1}}{h}\nabla\xi = 0. \tag{3.10}$$

It is convenient to define the "discretized velocity":

$$v^{h,j} := \frac{u^{h,j} - u^{h,j-1}}{h}.$$

We now want to derive an a priori estimate for the discrete energy:

$$E_j := E^{h,j} := \int_\Omega \phi(\nabla u^{h,j}, c^{h,j-1}) + \eta\psi(\nabla c^{h,j}, u^{h,j-1}) + \frac{1}{2}|v^{h,j}|^2\,dx,$$

where $\eta > 0$ will be chosen later.

We formulate the following lemma.

LEMMA 3.3 (discrete energy). *Let $T > 0$, $jh < T$, and $\delta \in (0, \varepsilon)$. Then for every positive $h < h_0(\delta)$ the following inequality holds:*

$$E_j + \sum_{j=1}^{\infty} \frac{h}{2} \int_\Omega (\varepsilon - \delta)|\nabla v^{h,j}|^2 dx + \sum_{j=1}^{\infty} \frac{h}{2} \int_\Omega (\varepsilon - \delta) \left| \frac{\nabla c^{h,j} - \nabla c^{h,j-1}}{h} \right|^2 dx \le M.$$

To prove this we exploit the fact that the nonconvex energy densities ϕ and ψ are "convexified" by the viscosity term. We start by considering the energy difference in one time step:

$$\begin{aligned}
\Delta E_j &:= E_{j+1} - E_j \\
&= \int_\Omega \left(\phi(\nabla u^{j+1}, c^j) + \frac{1}{2}|v^{j+1}|^2 + \eta\psi(\nabla c^{j+1}, u^j) \right) \\
&\quad - \left(\phi(\nabla u^j, c^{j-1}) + \frac{1}{2}|v^j|^2 + \eta\psi(\nabla c^j, u^{j-1}) \right) dx \\
&= \int_\Omega \Bigg(\phi(\nabla u^{j+1}, c^j) - \phi(\nabla u^j, c^{j-1}) \\
&\qquad + \frac{\delta}{h}|\nabla u^{j+1} - \nabla u^j|^2 - \frac{\delta}{h}|\nabla u^{j+1} - \nabla u^j|^2 - \frac{\delta}{h}\underbrace{|\nabla u^j - \nabla u^j|^2}_{=0} \\
&\qquad + \eta\psi(\nabla c^{j+1}, u^j) - \eta\psi(\nabla c^j, u^{j-1}) \\
&\qquad + \eta\frac{\delta}{h}|\nabla c^{j+1} - \nabla c^j|^2 - \eta\frac{\delta}{h}|\nabla c^{j+1} - \nabla c^j|^2 - \eta\frac{\delta}{h}\underbrace{|\nabla c^j - \nabla c^j|^2}_{=0} \\
&\qquad + \frac{1}{2}|v^{j+1}|^2 - \frac{1}{2}|v^j|^2 \Bigg) dx.
\end{aligned}$$

Before we proceed by estimating this expression, we first state the following auxiliary lemma.

LEMMA 3.4. *Let $r, s \ge 1$ and $\omega \in \mathcal{C}^2(\mathbb{R}^{r\times s}, \mathbb{R}_+)$. Assume that either ω satisfies, for every F_1, $F_2 \in \mathbb{R}^{r\times s}$, the inequality*

$$(\nabla\omega(F_1) - \nabla\omega(F_2))(F_1 - F_2) \ge -M|F_1 - F_2|^2, \tag{3.11}$$

where $M > 0$ is a constant, or that $\nabla\omega$ satisfies the Andrews–Ball condition; see (2.3).

Then for every $A \in \mathbb{R}^{r\times s}$ the function

$$g : F \longmapsto \omega(F) + \frac{\delta}{h}|F - A|^2$$

is convex for every $h \le h_0(\delta)$.

Furthermore for every F_1, $F_2 \in \mathbb{R}^{r\times s}$ and $h \le h_0(\delta)$ the following estimate holds:

$$\begin{aligned}&\left(\omega(F_1)+\frac{\delta}{h}|F_1-A|^2\right)-\left(\omega(F_2)+\frac{\delta}{h}|F_2-A|^2\right)\\ &\le\left(\nabla\omega(F_1)+\frac{2\delta}{h}(F_1-A)\right)(F_1-F_2).\end{aligned}\tag{3.12}$$

To prove the convexity of g we apply (3.11), which itself is a consequence of the Andrews–Ball condition (for a proof see, e.g., [11]). By the convexity of g we get $g(F_1)-g(F_2)\le \nabla g(F_1)(F_1-F_2)$, and this gives (3.12). ∎

We apply this lemma twice: once with $F_1 := \nabla u^{j+1}$, $F_2 := \nabla u^j$, $A := \nabla u^j$, and $\omega(X) := \phi(X, c^j)$ and once with $F_1 := \nabla c^{j+1}$, $F_2 := \nabla c^j$, $A := \nabla c^j$, and $\omega(X) := \psi(X, u^j)$. Furthermore we use the global Lipschitz continuity of S and K in the second variable (with Lipschitz constant L) to derive

$$\begin{aligned}\Delta E_j \le{}&\int_\Omega\left(\nabla_1\phi(\nabla u^{j+1},c^j)+\frac{2\delta}{h}(\nabla u^{j+1}-\nabla u^j)\right)(\nabla u^{j+1}-\nabla u^j)\\ &-\frac{\delta}{h}|\nabla u^{j+1}-\nabla u^j|^2+L|c^j-c^{j-1}|^2\\ &+\eta\left(\nabla_1\psi(\nabla c^{j+1},u^j)+\frac{2\delta}{h}(\nabla c^{j+1}-\nabla c^j)\right)(\nabla c^{j+1}-\nabla c^j)\\ &-\frac{\delta\eta}{h}|\nabla c^{j+1}-\nabla c^j|^2+L\eta|u^j-u^{j-1}|^2\\ &+\frac{1}{2}\left(|v^{j+1}|^2-|v^j|^2\right)dx.\end{aligned}$$

By rearranging the terms we get

$$\begin{aligned}\Delta E_j \le{}&\int_\Omega\left(\nabla_1\phi(\nabla u^{j+1},c^j)+\frac{\varepsilon}{h}(\nabla u^{j+1}-\nabla u^j)\right)(\nabla u^{j+1}-\nabla u^j)\\ &-\frac{\varepsilon-\delta}{h}|\nabla u^{j+1}-\nabla u^j|^2+L|c^j-c^{j-1}|^2\\ &+\eta\left(\nabla_1\psi(\nabla c^{j+1},u^j)+\frac{\varepsilon}{h}(\nabla c^{j+1}-\nabla c^j)\right)(\nabla c^{j+1}-\nabla c^j)\\ &-\eta\frac{\varepsilon-\delta}{h}|\nabla c^{j+1}-\nabla c^j|^2+L\eta|u^j-u^{j-1}|^2\\ &+\frac{1}{2}\left(|v^{j+1}|^2-|v^j|^2\right)dx.\end{aligned}\tag{3.13}$$

Before we continue with our estimate we now consider (3.9) with $\zeta := u^{j+1}-u^j$ (or to be precise a smooth sequence ζ_k converging to $u^{j+1}-u^j$ and considering the limit $k\to\infty$), which gives us the following expression:

$$\begin{aligned}&\int_\Omega S(\nabla u^{j+1},c^j)(\nabla u^{j+1}-\nabla u^j)\,dx\\ &=\int_\Omega -(v^{j+1}-v^j)v^{j+1}-\frac{\varepsilon}{h}|\nabla u^{j+1}-\nabla u^j|^2\,dx.\end{aligned}$$

Using the same ideas for (3.10) we get

$$\begin{aligned}&\int_\Omega K(\nabla c^{j+1},u^j)(\nabla c^{j+1}-\nabla c^j)\,dx\\ &=\int_\Omega -\frac{1}{h}|c^{j+1}-c^j|^2-\frac{\varepsilon}{h}|\nabla c^{j+1}-\nabla c^j|^2\,dx.\end{aligned}$$

We insert these equations into (3.13) and use the Poincaré inequality for $u^{j+1}-u^j$, i.e.,

$$-\int_\Omega \frac{\varepsilon-\delta}{2h}|\nabla u^{j+1}-\nabla u^j|^2 \le -M_P\int_\Omega \frac{\varepsilon-\delta}{2h}|u^{j+1}-u^j|^2,$$

to get the following estimate:

$$\begin{aligned}
E_{j+1}-E_j \le \int_\Omega & -(v^{j+1}-v^j)v^{j+1}+\frac{1}{2}\left(|v^{j+1}|^2-|v^j|^2\right)\\
&-\frac{\varepsilon-\delta}{h}|\nabla u^{j+1}-\nabla u^j|^2-\frac{\varepsilon}{h}|\nabla u^{j+1}-\nabla u^j|^2+\frac{\varepsilon}{h}|\nabla u^{j+1}-\nabla u^j|^2\\
&-\frac{\varepsilon-\delta}{h}|\nabla c^{j+1}-\nabla c^j|^2-\frac{\varepsilon\eta}{h}|\nabla c^{j+1}-\nabla c^j|^2+\frac{\varepsilon\eta}{h}|\nabla c^{j+1}-\nabla c^j|^2\\
&+L\eta|u^j-u^{j-1}|^2-\frac{\eta}{h}|c^{j+1}-c^j|^2+L|c^j-c^{j-1}|^2\,dx\\
\le \int_\Omega & -\frac{1}{2}|v^{j+1}+v^j|^2-\frac{\varepsilon-\delta}{2h}|\nabla u^{j+1}-\nabla u^j|^2-\eta\frac{\varepsilon-\delta}{h}|\nabla c^{j+1}-\nabla c^j|^2\\
&-\left(\frac{\varepsilon-\delta}{2h}M_P-L\eta\right)|u^{j+1}-u^j|^2+L\eta\left(|u^j-u^{j-1}|^2-|u^{j+1}-u^j|^2\right)\\
&-\frac{\eta}{h}|c^{j+1}-c^j|^2+L|c^j-c^{j-1}|^2\,dx.
\end{aligned}$$

If we choose $\eta \le \frac{\varepsilon-\delta}{2}\frac{M_P}{L}$, $h<\min\left(\frac{\eta}{L},1\right)$ and sum over all $j\ge 1$, then we get

$$E_j-E_0 \le -\sum_{i=1}^{j}(\varepsilon-\delta)\frac{h}{2}||\nabla v^i||^2-\sum_{i=1}^{j}(\varepsilon-\delta)\frac{h}{2}\left|\left|\frac{\nabla c^i-\nabla c^{i-1}}{h}\right|\right|^2+M(c_0,z_0).$$

This gives the statement of the lemma. □

The following inequality is an easy corollary of Lemma 3.3.

COROLLARY 3.5. *For every $T=kh>0$ and $h\le h_0(\varepsilon)$ there exists a constant $M>0$ such that*

$$\sup_j\left(||\nabla u^{h,j}||^2+||\nabla c^{h,j}||^2+||v^{h,j}||^2\right)+\varepsilon\sum_{j=1}^{\frac{T}{h}}h||\nabla v^{h,j}||^2+\varepsilon\sum_{j=1}^{\frac{T}{h}}h\left|\left|\frac{\nabla c^i-\nabla c^{i-1}}{h}\right|\right|^2 \le M<\infty.$$

To obtain the proof one simply applies the growth conditions for ϕ and ψ and Lemma 3.3. □

We are now able to prove the existence of solutions $(u^{h,j},c^{h,j})$ of our time-discretized system.

We first solve the time-step problem with the help of a variational ansatz; i.e., we consider for $u\in H_0^1(\Omega)$, $c\in H^1(\Omega)$ the functional

$$\begin{aligned}
W^{h,j}(u,c) := \int_\Omega & \phi(\nabla u,c^{h,j-1})+\psi(\nabla c,u^{h,j-1})\\
&+\frac{\varepsilon}{2h}|\nabla u-\nabla u^{h,j-1}|^2+\frac{1}{2h^2}|u-2u^{h,j-1}+u^{h,j-2}|^2\\
&+\frac{\varepsilon}{2h}|\nabla c-\nabla c^{h,j-1}|^2+\frac{1}{2h}|c-c^{h,j-1}|^2\,dx.
\end{aligned}$$

The functional $W^{h,j}$ is weakly lower semicontinuous since its integrand is convex in $(\nabla u, \nabla c)$, which is true since the "critical" terms $\phi(\nabla u, c^{h,j-1}) + \frac{1}{2h}|\nabla u|^2$ and $\psi(\nabla c, u^{h,j-1}) + \frac{1}{2h}|\nabla c|^2$ are convex for sufficiently small $h > 0$.

Since $W^{h,j}$ is also bounded from below by zero, there exists a (not necessarily unique) minimizer (u, c). By a standard calculation one can show that (u, c) solves the time-step problem. We define $(u^{h,j}, c^{h,j}) := (u, c)$. By induction we get the existence of a time-discretized solution to the discrete problem.

In the next step we interpolate this discrete approximation $(u^{h,j}, c^{h,j})$ in time. Here it is convenient to use two different approximation schemes, i.e., the piecewise constant and the piecewise affine interpolation.

We define for $h > 0$, $0 \le j < \frac{T}{h}$, and the characteristic function $\chi^{h,j} := \chi_{[hj,h(j+1)]}$:

- $w^h(t) := \sum_j \chi^{h,j}(t) \frac{v^{h,j+1} - v^{h,j}}{h}$ (step function appr. of u_{tt}),
  $\tilde{v}^h(t) := \sum_j \chi^{h,j}(t) \left(v^{h,j} + \frac{v^{h,j+1} - v^{h,j}}{h}(t - hj) \right)$ (its primitive),
- $v^h(t) := \sum_j \chi^{h,j}(t) v^{h,j+1}$ (step function appr. of u_t),
  $\tilde{u}^h(t) := \sum_j \chi^{h,j}(t) \left(u^{h,j} + v^{h,j+1}(t - hj) \right)$ (its primitive),
- $u^h(t) := \sum_j \chi^{h,j}(t) u^{h,j+1}$ (step function appr. of u),
- $d^h(t) := \sum_j \chi^{h,j}(t) \frac{c^{h,j+1} - c^{h,j}}{h}$ (step function appr. of c_t),
  $\tilde{c}^h(t) := \sum_j \chi^{h,j}(t) \left(c^{h,j} + \frac{c^{h,j+1} - c^{h,j}}{h}(t - hj) \right)$ (its primitive),
- $c^h(t) := \sum_j \chi^{h,j}(t) c^{h,j+1}$ (step function appr. of c).

We have chosen the notation in such a way that the step functions are each labeled with different characters (w, v, u, resp., d and c) depending on the order of derivative they are approximating. Their primitives are denoted by the character of the corresponding lower order terms with a tilde; e.g., the primitive of w^h is denoted as $\tilde{v}^h$. Later we will show that the interpolations of the same character with or without a tilde (i.e., terms of the same order) coincide in the limit $h \to 0$ and converge to our solution or its derivatives.

To prove convergence for these sequences we use Corollary 3.5, and we use the growth conditions (in the cases where the H^{-1}-norm is involved we also use the discretized partial differential equations) to prove the following bounds (uniformly in h) for fixed $T > 0$:

$$\begin{aligned}
\sup_{0 \le t \le T} ||u^h(t)||^2_{H^1_0} &\le M(u_0, z_0, c_0), \\
\sup_{0 \le t \le T} ||\tilde{u}^h(t)||^2_{H^1_0} &\le M(u_0, z_0, c_0), \\
\sup_{0 \le t \le T} ||v^h(t)||^2 &\le M(u_0, z_0, c_0), \\
\varepsilon \int_0^T ||v^h(t)||^2_{H^1_0} \, dt &\le M(u_0, z_0, c_0), \\
\sup_{0 \le t \le T} ||\tilde{v}^h(t)||^2 &\le M(u_0, z_0, c_0),
\end{aligned}$$

$$\begin{aligned}
\varepsilon \int_0^T ||\tilde{v}^h(t)||^2_{H^1_0}\,dt &\le M(u_0,z_0,c_0),\\
\sup_{0\le t\le T} ||w^h(t)||^2_{H^{-1}} &\le M(u_0,z_0,c_0),\\
\sup_{0\le t\le T} ||c^h(t)||^2_{H^1} &\le M(u_0,z_0,c_0),\\
\sup_{0\le t\le T} \left(||\tilde{c}^h(t)||^2 + ||d^h(t)||^2_{H^{-1}}\right) &\le M(u_0,z_0,c_0),\\
\varepsilon \int_0^T ||d^h(t)||^2_{H^1}\,dt &\le M(u_0,z_0,c_0).
\end{aligned}$$

From these bounds we get the following weak convergence results (again choosing subsequences) for $h \to 0$:

$$\begin{aligned}
&u^h \overset{\star}{\rightharpoonup} u \quad \text{in } L^\infty\left((0,T),H^1(\Omega)\right),\\
&\tilde{u}^h \overset{\star}{\rightharpoonup} \tilde{u} \quad \text{in } L^\infty\left((0,T),H^1(\Omega)\right)\cap W^{1,\infty}\left((0,T),L^2(\Omega)\right)\cap W^{1,2}\left((0,T),H^1(\Omega)\right),\\
&v^h \overset{\star}{\rightharpoonup} v \quad \text{in } L^\infty\left((0,T),L^2(\Omega)\right)\cap L^2\left((0,T),H^1(\Omega)\right),\\
&\tilde{v}^h \overset{\star}{\rightharpoonup} \tilde{v} \quad \text{in } L^\infty\left((0,T),L^2(\Omega)\right)\cap L^2\left((0,T),H^1(\Omega)\right)\cap W^{1,\infty}\left((0,T),H^{-1}(\Omega)\right),\\
&w^h \overset{\star}{\rightharpoonup} w \quad \text{in } L^2\left((0,T),H^{-1}(\Omega)\right),\\
&c^h \overset{\star}{\rightharpoonup} c \quad \text{in } L^\infty\left((0,T),H^1(\Omega)\right),\\
&\tilde{c}^h \overset{\star}{\rightharpoonup} \tilde{c} \quad \text{in } W^{1,\infty}\left((0,T),H^{-1}(\Omega)\right),\\
&d^h \overset{\star}{\rightharpoonup} d \quad \text{in } L^2\left((0,T),H^1(\Omega)\right).
\end{aligned}$$

Additionally we deduce by applying Corollary 3.5 and the growth conditions on S and K that there exists $\tilde{S}$ and $\tilde{K}$ such that for $\hat{c} \in L^\infty((0,T),L^2(\Omega,\mathbb{R}^d))$, $\hat{u} \in L^\infty((0,T),L^2(\Omega,\mathbb{R}^m))$,

$$\begin{aligned}
\sup_{0\le t\le T}\int_\Omega |S(\nabla u^h,\hat{c})|^2 &\le M \sup_{0\le t\le T}(||\nabla u^h||^2 + ||\hat{c}||^2 + 1) \;\le\; M(\hat{c}),\\
\sup_{0\le t\le T}\int_\Omega |K(\nabla c^h,\hat{u})|^2 &\le M \sup_{0\le t\le T}(||\nabla c^h||^2 + ||\hat{u}||^2 + 1) \;\le\; M(\hat{u}),
\end{aligned}$$

and hence (for subsequences)

$$\begin{aligned}
S(\nabla u^h,\hat{c}) &\overset{\star}{\rightharpoonup} \tilde{S}_c \quad \text{in } L^\infty\left((0,T),L^2(\Omega)\right),\\
K(\nabla c^h,\hat{u}) &\overset{\star}{\rightharpoonup} \tilde{K}_u \quad \text{in } L^\infty\left((0,T),L^2(\Omega)\right).
\end{aligned}$$

We now have to make sure that the different interpolations we have chosen converge to the same limit. For this we use a standard lemma (see, e.g., [12]).

LEMMA 3.6. *Suppose that $(f^{h,j})_{h,j}$ is bounded in $L^2(\Omega)$, that $f^h(t)$ is its step function interpolation, and that $g^h(t)$ is its continuous and piecewise affine interpolation. Assume furthermore that $f^h \rightharpoonup f$ and $g^h \rightharpoonup g$ in $L^2_{loc}(\Omega\times\mathbb{R}^+)$. Then we have $f = g$.*

Sketch of the proof. We show the equivalence after testing with a smooth function. Therefore we need only consider test functions of the "separated" form $w(x)z(t)$. Let $\xi^h(t)$ be the step function approximation of $z(t)$ and let $\zeta^h(t)$ be the piecewise affine

approximation of $z(t)$. Then $w(x)\xi^h(t)$ and $w(x)\zeta^h(t)$ converge strongly to $w(x)z(t)$. If we now test $f^h(t)$ with $w(x)\xi^h(t)$ and $g^h(t)$ with $w(x)\zeta^h(t)$, we get the same result, and this equation holds also for $h \to 0$. (See [12] for the complete proof.) ∎

We can apply this lemma to deduce $u = \tilde{u}$, $v = \tilde{v}$, and $c = \tilde{c}$. This is nearly enough to consider the limit $h \to 0$ in our equation, but the nonlinearities S and K cannot be handled in this way, since weak convergence of ∇u^h to ∇u is not enough to get weak convergence of $S(\nabla u^h, c)$ to $S(\nabla u, c)$. (And the analogous statement holds for K.) Fortunately we can prove strong convergence of ∇u^h, $\nabla \tilde{u}^h$, ∇c^h, and $\nabla \tilde{c}^h$ in $L^2\left((0,T), L^2(\Omega)\right)$ as $h \to 0$.

We first need some lemmas, where we state only simplified counterparts of the corresponding lemmas in [11]. The proofs can also be found there.

LEMMA 3.7 (Aubin-type result). *Let $X_s := W^{1,2}(\Omega)$, $X := L^2(\Omega)$, and $X_w := W^{-1,2}(\Omega)$. Then the imbedding of $L^2((0,T), X_s) \cap W^{1,2}((0,T), X_w)$ equipped with the natural norm $||\cdot||_{L^2(X_s)} + ||\partial_t \cdot||_{L^2(X_w)}$ into $L^2((0,T), X)$ is compact.*

The next lemma gives a closer connection between the two kinds of interpolations we have used.

LEMMA 3.8. *Let X be a Banach space and $\{f^{h,j}\}_{j\geq 1, h>0}$ a collection of elements in X. Let f^h be the piecewise constant and let $\tilde{f}^h$ be the piecewise linear interpolation of $\{f^{h,j}\}$ defined (as usual) by*

$$f^h(t) := \sum_j \chi_j(t) f^{h,j},$$

$$\tilde{f}^h(t) := \sum_j \chi_j(t) \left(\left(j - \frac{t}{h} \right) f^{h,j-1} + \left(\frac{t}{h} - (j-1) \right) f^{h,j} \right),$$

where χ_j is the characteristic function of $(jh, (j+1)h)$.

Assume that $\sup_j ||f^{h,j}||^2 \leq M_1$ and for some $\alpha > 0$,

$$\sum_{j=1}^{\frac{T}{h}} h \left|\left| \frac{f^{h,j} - f^{h,j-1}}{h^\alpha} \right|\right|^2 \leq M_2.$$

Then for all $f \in L^2((0,T), X)$ with $\sup_t ||f(t)||^2 \leq M_1$ we have the following estimate:

$$\int_0^T ||f^h - f||^2 \, dt \leq 2 \int_0^T ||\tilde{f}^h - f||^2 \, dt + 4hM_1 + \frac{2}{3} h^{2\alpha} M_2.$$

We also use the following fact following from the definition of weak convergence and compactness.

LEMMA 3.9. *Let $G \subset \mathbb{R}^N$ be open, let $\{f^h\}_h \subset L^2(\Omega)$, let $f^h \rightharpoonup 0$ in $L^2(\Omega)$ as $h \to 0$, and let K be a compact subset of $L^2(\Omega)$; then*

$$\sup_{\xi \in K} \left| \int_G f^h \xi \, dx \right| \to 0 \quad \text{as } h \to 0.$$

Now we have collected all ingredients for the proof of the strong convergence of ∇u^h, $\nabla \tilde{u}^h$, ∇c^h, and $\nabla \tilde{c}^h$. First we consider ∇u^h and $\nabla \tilde{u}^h$, and later we will apply the methods introduced there to prove the strong convergence of ∇c^h and $\nabla \tilde{c}^h$.

We start with the following time-integrated version of our elasticity equation, which does not require that the test function ζ be differentiable in time; i.e., for

$\zeta \in L^2((0,T), H_0^1(\Omega))$,

$$\int_0^T \int_\Omega \left(S(\nabla u^h, c^h(\cdot - h)) + \varepsilon \nabla v^h\right) \nabla \zeta - v^h \frac{\zeta(\cdot + h) - \zeta}{h} \, dx \, dt$$
$$+ \frac{1}{h} \int_{T-h}^T \int_\Omega v^h \zeta(\cdot + h) \, dx \, dt - \frac{1}{h} \int_{-h}^0 \int_\Omega v_0 \zeta(\cdot + h) \, dx \, dt = 0. \tag{3.14}$$

We consider the limit $h \to 0$ in equation (3.14), where we use that $c^h(\cdot - h) \to c$ in $L^\infty((0,T), L^2(\Omega))$. Using the definition of $\tilde{S}_c$ we get

$$\int_0^T \int_\Omega \tilde{S}_c \nabla \zeta + \varepsilon \nabla u_t \nabla \zeta - u_t \zeta_t \, dx \, dt$$
$$+ \int_\Omega \underbrace{v(T)\zeta(T) + v_0 \zeta(0)}_{=0} \, dx = 0. \tag{3.15}$$

We insert $\zeta := u^h - u$ in (3.14) and $\zeta := \tilde{u}^h - u$ in (3.15) and substract the resulting equations. (To be exact we have to approximate $u^h - u^h(\cdot - h)$ and $\tilde{u}^h - u$ by sequences of smooth functions.) This gives, for $t \in (0,T)$,

$$0 = \underbrace{\int_0^t \int_\Omega S(\nabla u^h, c^h(\cdot - h))(\nabla u^h - \nabla u) - \tilde{S}_c(\nabla \tilde{u}^h - \nabla u) \, dx \, dt}_{=:T_1}$$
$$+ \varepsilon \underbrace{\int_0^t \int_\Omega \nabla v^h (\nabla u^h - \nabla u) - \nabla u_t (\nabla \tilde{u}^h - \nabla u) \, dx \, dt}_{=:T_2}$$
$$\underbrace{- \int_0^t \int_\Omega v^h \left(v^h(\cdot + h) - \frac{\tilde{u}(\cdot + h) - u}{h} \right) - u_t \left((\tilde{u}^h)_t - u_t \right) \, dx \, dt}_{=:T_3}$$
$$\underbrace{+ \int_\Omega \int_{t-h}^t v^h (u^h(\cdot + h) - u(\cdot + h)) - u_t \left(\tilde{u}^h - u \right) \, dt \, dx}_{=:T_4}$$
$$\underbrace{- \int_\Omega v_0 \frac{1}{h} \int_{-h}^0 u^h(\cdot + h) - u(\cdot + h) \, dt \, dx}_{=:T_5},$$

where we have defined the terms $T_1, \ldots, T_5$, which we will estimate in the following calculation. To simplify notation we denote all terms converging to zero as $h \to 0$ (uniformly in t) by $\alpha(h)$.

We start by estimating T_1, where we use the global Lipschitz continuity of S, giving us for a certain $M > 0$ and every $F_1, F_2 \in \mathbb{R}^{m \times n}$ and $\hat{c} \in \mathbb{R}^d$ the inequality

$$(S(F_1, \hat{c}) - S(F_2, \hat{c})(F_1 - F_2)) \geq -M |F_1 - F_2|^2.$$

(This corresponds to condition (2.4) in the last section.)

$$T_1 = \int_0^t \int_\Omega \left(S(\nabla u^h, c^h(\cdot - h)) - S(\nabla u, c^h(\cdot - h)) \right) (\nabla u^h - \nabla u)$$

$$+ S(\nabla u, c^h(\cdot - h))(\nabla u^h - \nabla u) - \tilde{S}_c(\nabla \tilde{u}^h - \nabla u)\,dx\,dt$$
$$\geq -M \int_0^t \int_\Omega |\nabla u^h - \nabla u|^2\,dx\,dt - \sup_{t\in(0,T)} \left| \int_0^T \int_\Omega \left(\chi_{\Omega\times(0,t)} \tilde{S}_c\right)(\nabla \tilde{u}^h - \nabla u)\,dx\,dt \right|^2.$$

Applying Lemma 3.9 we can show that the last three terms converge to zero for $h \to 0$, i.e.,

$$T_1 \geq -M \int_0^t \int_\Omega |\nabla u^h - \nabla u|^2 + \alpha(h).$$

Applying Lemma 3.8 we finally get

$$T_1 \geq -2M \int_0^t \int_\Omega |\nabla \tilde{u}^h - \nabla u|^2 + \alpha(h).$$

We can use the same calculations as in the purely viscoelastic case (see [11][1]) to derive

$$-T_2 = -\frac{1}{2} \int_\Omega |\nabla \tilde{u}^h(t) - \nabla u(t)|^2\,dx + \frac{1}{2} \int_\Omega \underbrace{|\nabla \tilde{u}^h(0) - \nabla u(0)|^2}_{=0}\,dx + \alpha(h),$$

where the discrete energy estimate proved above is used. This (together with the estimate for T_1) is the key step to the desired strong convergence result, since at the end we want to apply the Gronwall lemma to the inequality we get by estimating these terms. Therefore we need the terms T_3–T_5 to be "well behaved," i.e., that they are simply $\alpha(h)$.

In fact by applying Lemma 3.7 combined with Lemma 3.8 we can prove this:

$$T_3 = \alpha(h), \quad T_4 = \alpha(h), \quad T_5 = \alpha(h).$$

Taking everything together we have the inequality

$$\partial_t \int_0^t \int_\Omega |\nabla \tilde{u}^h - \nabla u|^2\,dx\,dt \leq \frac{4M}{\varepsilon} \int_0^t \int_\Omega |\nabla \tilde{u}^h - \nabla u|^2\,dx\,dt + \alpha(h).$$

Now we can apply the Gronwall lemma to get

$$\int_0^T \int_\Omega |\nabla \tilde{u}^h - \nabla u|^2\,dx\,dt \leq \alpha(h) \frac{\varepsilon}{4M} e^{\frac{4MT}{\varepsilon}},$$

and this converges to zero for $h \to 0$. Hence

$$\nabla \tilde{u}^h \to \nabla u \quad \text{in } L^2\left((0,T), L^2(\Omega)\right).$$

Due to Lemma 3.8 the same convergence result holds for ∇u^h. This ensures that $\tilde{S}_{\hat{c}} = S(\nabla u, \hat{c})$.

Now we are ready to apply the same methods to prove $\tilde{K}_{\hat{u}} = K(\nabla c, \hat{u})$. First we consider the following weak formulation of (3.10):

$$\int_0^T \int_\Omega K(\nabla c^h, u^h(\cdot - h))\nabla\zeta + \varepsilon \frac{\nabla c^h - \nabla c^h(\cdot - h)}{h} \nabla\zeta + \frac{c^h - c^h(\cdot - h)}{h} \zeta\,dx\,dt = 0. \tag{3.16}$$

[1]Recall the slightly different notation in their article.

Then we consider the limit $h \to 0$ in (3.16) to get

$$\int_0^T \int_\Omega \tilde{K}_u \nabla\zeta + \varepsilon \nabla c_t \nabla\zeta + c_t \zeta \, dx \, dt = 0. \tag{3.17}$$

We insert $\zeta := c^h - c$ in (3.16) and $\zeta := \tilde{c}^h - c$ in (3.17) and substract the resulting equations. (To be exact we have to approximate $c^h - c^h(\cdot - h)$ and $\tilde{c}^h - c$ by sequences of smooth functions.) This gives

$$\begin{aligned} 0 = \int_0^T \int_\Omega & K(\nabla c^h, u^h(\cdot - h)) \nabla(c^h - c) - \tilde{K}_u \nabla(\tilde{c}^h - c) \\ & + \varepsilon \frac{\nabla c^h - \nabla c^h(\cdot - h)}{h} \nabla(c^h - c) - \varepsilon \nabla c_t \nabla(\tilde{c}^h - c) \\ & + \frac{c^h - c^h(\cdot - h)}{h} (c^h - c) - c_t(\tilde{c}^h - c) \, dx \, dt. \end{aligned} \tag{3.18}$$

Now we consider the three terms in (3.18). We start with the third one. We want to prove that

$$\int_0^T \int_\Omega \frac{c^h - c^h(\cdot - h)}{h} (c^h - c) - c_t(\tilde{c}^h - c) \, dx \, dt \to 0 \quad \text{as } h \to 0.$$

But this is true for the first part, since $\left(\tilde{c}^h\right)_t$ is bounded in $L^\infty((0,T), H^{-1}(\Omega))$ and $c^h \overset{\star}{\rightharpoonup} c$ in $L^\infty((0,T), H^1(\Omega))$, and it is true for the second part, since $c_t \in L^2((0,T), H^1(\Omega))$ and $\tilde{c}^h \overset{\star}{\rightharpoonup} c$ in $W^{1,\infty}(H^{-1}(\Omega))$.

We now rewrite the first term in (3.18), denote it by T_6, and estimate it as follows:

$$\begin{aligned} T_6 := \int_0^t \int_\Omega & \left(K(\nabla c^h, u^h(\cdot - h)) - K(\nabla c, u^h(\cdot - h))\right) \nabla(c^h - c) \\ & + K(\nabla c, u^h(\cdot - h)) \nabla(c^h - c) \\ & - \tilde{K}_{u^h(\cdot - h)} \nabla(\tilde{c}^h - c) + \left(\tilde{K}_{u^h(\cdot - h)} - \tilde{K}_u\right) \nabla(\tilde{c}^h - c) \, dx \, dt \\ \geq & -M \int_0^t \int_\Omega |\nabla \tilde{c}^h - \nabla c|^2 \, dx \, dt + \alpha(h). \end{aligned}$$

So we get

$$T_6 \geq -M \int_0^t \int_\Omega |\nabla \tilde{c}^h - \nabla c|^2 \, dx \, dt + \alpha(h). \tag{3.19}$$

It remains to estimate the second term in (3.18). Here we apply the methods we had used to estimate T_2. This gives the following inequality:

$$\begin{aligned} & -\varepsilon \int_0^T \int_\Omega \frac{\nabla c^h - \nabla c^h(\cdot - h)}{h} (\nabla c^h - \nabla c) - \nabla c_t (\nabla \tilde{c}^h - \nabla c) \\ & \geq -\frac{\varepsilon}{2} \int_0^T \int_\Omega |\nabla \tilde{c}^h(t) - \nabla u(t)|^2 \, dx \, dt + \alpha(h). \end{aligned} \tag{3.20}$$

If we insert (3.19) and (3.20) into (3.18) and apply the Gronwall lemma in the same way as before, we derive

$$\int_0^T \int_\Omega |\nabla \tilde{c}^h - \nabla c|^2 \, dx \, dt \leq \alpha(h) \frac{\varepsilon}{4M} e^{\frac{4MT}{\varepsilon}},$$

and this is converging to zero for $h \to 0$. Therefore $\nabla \tilde{c}^h$ is converging to ∇c strongly in $L^2((0,T), L^2(\Omega))$. And due to Lemma 3.8 this also holds for ∇c^h. Hence for $h \to 0$ the nonlinear term $K(\nabla c^h, u^h)$ converges to $K(\nabla c, u)$.

Taking everything together we have proved that the solutions of the time-discretized equations converge to solutions of the hyperbolic-parabolic system (3.5).

To prove Theorem 3.2 it remains only to prove the energy inequality,

$$E(t) \le M(u_0, v_0, c_0) - \varepsilon \int_0^t \int_\Omega |\nabla u_t|^2 + |\nabla c_t|^2 \, dx \, dt,$$

where $E(t) := \int_\Omega \phi(\nabla u(t), c(t)) + \psi(\nabla c(t), u(t)) + \frac{1}{2}|u_t(t)|^2 \, dx$ and $M(u_0, v_0, c_0)$ is a constant depending only on the initial values u_0, v_0, c_0.

To prove this we start from the discrete energy inequality (Lemma 3.3), telling us that for $\eta > 0$ sufficiently small, $h < \min\left(1, \frac{\eta}{L}\right)$, and $\delta \in (0,1)$ the following inequality holds for every $t \in (0,T)$:

$$\begin{aligned} &\int_\Omega \phi(\nabla u^h, c^h(\cdot - h)) \, dx + \eta \int_\Omega \psi(\nabla c^h, u^h(\cdot - h)) \, dx \\ &+ \frac{1}{2} \int_\Omega |v^h|^2 \, dx + (\varepsilon - \delta) \int_0^t \int_\Omega |\nabla v^h|^2 + |\nabla d^h|^2 \, dx \, dt \le M(u_0, v_0, c_0). \end{aligned}$$

Now we notice that we can apply these convergence results:

$$\begin{aligned} v^h(t) &\to u_t(t) \quad \text{in } L^2(\Omega) \text{ for almost every } t \in (0,T), \\ \nabla v^h &\rightharpoonup \nabla u_t \quad \text{in } L^2((0,T) \times \Omega), \\ \nabla d^h &\rightharpoonup \nabla c_t \quad \text{in } L^2((0,T) \times \Omega). \end{aligned}$$

By the weakly lower semicontinuity of the $L^2((0,T)\times\Omega)$-norm we get for almost every $t \in (0,T)$

$$\begin{aligned} &\limsup_{h \to 0} \int_\Omega \phi(\nabla u^h, c^h(\cdot - h)) \, dx + \eta \limsup_{h \to 0} \int_\Omega \psi(\nabla c^h, u^h(\cdot - h)) \, dx \\ &+ \frac{1}{2} \int_\Omega |u_t|^2 \, dx + (\varepsilon - \delta) \int_0^t \int_\Omega |\nabla u_t|^2 + |\nabla c_t|^2 \, dx \, dt \le M(u_0, v_0, c_0). \end{aligned} \tag{3.21}$$

Now we apply the strong convergence of $\nabla u^h(t)$ to estimate

$$\begin{aligned} \int_\Omega \phi(\nabla u, c) \, dx &= \int_\Omega \left(\phi(\nabla u, c) + \frac{M}{2} |\nabla u|^2 \right) dx - \int_\Omega \frac{M}{2} |\nabla u|^2 \, dx \\ &\le \limsup_{h \to 0} \int_\Omega \left(\phi(\nabla u^h, c^h(\cdot - h)) + \frac{M}{2} |\nabla u^h|^2 \right) dx \\ &\quad - \int_\Omega \frac{M}{2} |\nabla u|^2 \, dx \\ &\le \limsup_{h \to 0} \int_\Omega \phi(\nabla u^h, c^h(\cdot - h)) \, dx + \limsup_{h \to 0} \int_\Omega \frac{M}{2} |\nabla u^h|^2 \, dx \\ &\quad - \int_\Omega \frac{M}{2} |\nabla u|^2 \, dx \\ &= \limsup_{h \to 0} \int_\Omega \phi(\nabla u^h, c^h(\cdot - h)) \, dx. \end{aligned}$$

Similarly, applying the strong convergence of $\nabla c^h(t)$ we get

$$\int_\Omega \psi(\nabla c, u)\,dx = \limsup_{h\to 0} \int_\Omega \psi(\nabla c^h, u^h(\cdot - h))\,dx.$$

If we insert these estimates into (3.21) and take the limit $\delta \to 0$, then we get

$$\int_\Omega \phi(\nabla u, c)\,dx + \eta \int_\Omega \psi(\nabla c, u)\,dx + \frac{1}{2}\int_\Omega |u_t|^2\,dx \\ + \varepsilon \int_0^t \int_\Omega |\nabla u_t|^2 + |\nabla c_t|^2\,dx\,dt \le M(u_0, v_0, c_0)$$

for almost every $t \in (0,T)$.

By adjusting the constant M we get the desired estimate (3.8). This completes the proof of Theorem 3.2. □

Now we apply this to prove Theorem 3.1 by considering $\varepsilon \to 0$ in the same spirit as in the previous section: First the energy inequality (3.8) gives the following weak convergence results (for subsequences) as $\varepsilon \to 0$:

$$u^\varepsilon \overset{\star}{\rightharpoonup} u \quad \text{in } L^\infty((0,T), H_0^1(\Omega)),$$
$$c^\varepsilon \overset{\star}{\rightharpoonup} c \quad \text{in } L^\infty((0,T), H_0^1(\Omega)),$$
$$u^\varepsilon \overset{\star}{\rightharpoonup} u \quad \text{in } W^{1,\infty}((0,T), L^2(\Omega)).$$

Furthermore for almost every $t \in (0,T)$ the sequence $\nabla u^\varepsilon(t)$ generates the Young measure $\nu_{\cdot,t}$ and $\nabla c^\varepsilon(t)$ generates the Young measure $\mu_{\cdot,t}$, since $||\nabla u^\varepsilon(t)||$ and $||\nabla c^\varepsilon(t)||$ are bounded (uniformly in t).

In particular we get that for any $\zeta \in H_0^1(\Omega)$

$$\int_\Omega S(\nabla u^\varepsilon(\cdot,t), c^\varepsilon(\cdot,t))\nabla\zeta \to \int_\Omega \langle \nu_{\cdot,t}, S(\cdot, c(\cdot,t))\rangle \nabla\zeta.$$

For a subsequence we can consider the limit of the viscoelastic equations for $\varepsilon \to 0$ by using the growth conditions on S and K and the strong convergence of u^ε and c^ε in L^2: Since for all $\varepsilon > 0$

$$||S(\nabla u^\varepsilon, c^\varepsilon)||_{L^\infty((0,T),L^2(\Omega))} \le M,$$

we obtain the existence of a function $\tilde{S} \in L^\infty((0,T), L^2(\Omega))$ with

$$\int_0^T \int_\Omega S(\nabla u^\varepsilon, c^\varepsilon)\nabla\zeta \to \int_0^T \int_\Omega \tilde{S}\nabla\zeta.$$

Taking both together and repeating the calculations from section 2 to estimate the other terms, we obtain (3.3) and (3.4). The Neumann boundary condition on c^ε is converging to the Neumann boundary condition on c. This calculation concludes the proof of existence for Theorem 3.1. □

An easy consequence of this theorem is the following corollary.

COROLLARY 3.10 (vector-valued parabolic equations). *Under the assumptions on ψ, K, and c_0 stated above there exists a YM-solution, defined in an analogous way to YM-solutions for hyperbolic-parabolic systems, for the parabolic system*

$$c_t(x,t) - \operatorname{div} K(\nabla c(x,t)) = 0,$$
$$c(\cdot,0) = c_0,$$
$$\vec{n} K(\nabla c, u) = 0 \quad \text{on } \partial\Omega,$$

where $c\colon \Omega \to \mathbb{R}^m$.

Proof. To prove this we only have to study the uncoupled case in Theorem 3.1, which is not excluded by the growth conditions. □

This result is an extension of a scalar version that can be found in [6].

4. Concluding remarks. The regularity of the YM-solutions as constructed in the last sections is still a widely open problem. Although there are easy examples for YM-solutions which are no weak solutions even in the one-dimensional parabolic case (see, e.g., [18]), to the author's knowledge, there is no example where *every* YM-solution to a given data fails to be a weak solution. It would be very interesting to find an example where a smooth initial data develops a microstructure in finite time. For some related results, see [22], [23].

Acknowledgments. I am grateful to Stefan Müller for his steady and valuable help, and to Sophia Demoulini and Johannes Zimmer for very stimulating discussions.

REFERENCES

[1] G. Andrews and J. M. Ball, *Asymptotic behaviour and changes of phase in one-dimensional nonlinear viscoelasticity*, J. Differential Equations, 44 (1982), pp. 306–341.
[2] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Ration. Mech. Anal., 63 (1977), pp. 337–403.
[3] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Ration. Mech. Anal., 100 (1987), pp. 13–52.
[4] C. Carstensen and M. O. Rieger, *Numerical Simulations in Non-Monotone Elastodynamics Involving Young-Measure Approximations*, Preprint 27/2002, Center for Nonlinear Analysis, Carnegie Mellon University, Pittsburgh, PA, 2002.
[5] C. M. Dafermos and W. J. Hrusa, *Energy methods for quasilinear hyperbolic initial-boundary value problems. Applications to elastodynamics*, Arch. Ration. Mech. Anal., 87 (1985), pp. 267–292.
[6] S. Demoulini, *Young measure solutions for a nonlinear parabolic equation of forward-backward type*, SIAM J. Math. Anal., 27 (1996), pp. 376–403.
[7] S. Demoulini, *Young measure solutions for nonlinear evolutionary systems of mixed type*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 143–162.
[8] S. Demoulini, *Weak solutions for a class of nonlinear systems of viscoelasticity*, Arch. Ration. Mech. Anal., 155 (2000), pp. 299–334.
[9] S. Demoulini, D. M. Stuart, and A. E. Tzavaras, *A variational approximation scheme for three-dimensional elastodynamics with polyconvex energy*, Arch. Ration. Mech. Anal., 157 (2001), pp. 325–344.
[10] R. J. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch. Ration. Mech. Anal., 82 (1983), pp. 27–70.
[11] G. Friesecke and G. Dolzmann, *Implicit time discretization and global existence for a quasilinear evolution equation with nonconvex energy*, SIAM J. Math. Anal., 28 (1997), pp. 363–380.
[12] D. Kinderlehrer and P. Pedregal, *Weak convergence of integrands and the Young measure representation*, SIAM J. Math. Anal., 23 (1992), pp. 1–19.
[13] D. Kinderlehrer and P. Pedregal, *Gradient Young measures generated by sequences in Sobolev spaces*, J. Geom. Anal., 4 (1994), pp. 59–90.
[14] J. Málek, J. Nečas, M. Rokyta, and M. Ružička, *Weak and Measure-Valued Solutions to Evolutionary PDEs*, Chapman and Hall, London, 1996.
[15] S. Müller, *Variational models for microstructure and phase transitions*, in Calculus of Variations and Geometric Evolution Problems, S. Hildebrandt and M. Struwe, eds., Lecture Notes in Math. 1713, Springer-Verlag, Berlin, 1999, pp. 85–210.
[16] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser-Verlag, Basel, 1997.
[17] M. O. Rieger, *Young-measure solutions for an elasticity equation with diffusion*, in EQUADIFF 99, International Conference on Differential Equations, World Scientific, River Edge, NJ, 2000, pp. 457–459.
[18] M. O. Rieger, *Nonconvex Dynamical Problems*, Ph.D. thesis, Max-Planck-Institute for Mathematics in the Sciences, Leipzig, Germany, 2001.

[19] J. W. Shearer, *Global existence and compactness in L^p for the quasi-linear wave equation*, Comm. Partial Differential Equations, 19 (1994), pp. 1829–1877.
[20] M. Slemrod, *Dynamics of measured valued solutions to a backward-forward heat equation*, J. Dynam. Differential Equations, 3 (1991), pp. 1–28.
[21] V. Šverák and J. Nečas, *Personal communication*, 2000.
[22] F. Theil, *Young-measure solutions for a viscoelastically damped wave equation with nonmonotone stress-strain relation*, Arch. Ration. Mech. Anal., 144 (1998), pp. 47–78.
[23] F. Theil and V. I. Levitas, *A study of a Hamiltonian model for martensitic phase transformations including microkinetic energy*, Math. Mech. Solids, 5 (2000), pp. 337–368.
[24] J. Zimmer, *Mathematische Modellierung und Analyse von Formgedächtnislegierungen in mehreren Raumdimensionen*, Ph.D. thesis, Technische Universität München, Germany, 2000.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1399–1421

ON THE OORT–HULST–SAFRONOV COAGULATION EQUATION AND ITS RELATION TO THE SMOLUCHOWSKI EQUATION*

MIROSŁAW LACHOWICZ†, PHILIPPE LAURENÇOT‡, AND DARIUSZ WRZOSEK†

Abstract. A connection is established between the classical Smoluchowski continuous coagulation equation and the Oort–Hulst–Safronov coagulation equation via generalized coagulation equations. Existence of solutions to the Oort–Hulst–Safronov coagulation equation is shown, and the large time behavior and the occurrence of gelation are studied as well. It is also shown that a compactly supported initial distribution propagates with finite speed.

Key words. coagulation, Smoluchowski equation, Oort–Hulst–Safronov equation, existence, gelation, finite speed of propagation

AMS subject classifications. 45K05, 45M05, 45G10, 82C21

PII. S0036141002414470

1. Introduction.

Coagulation processes are found in a wide variety of physical situations where clusters (or particles, droplets, etc.) merge by coalescence to form larger ones. Such a phenomenon takes place in, e.g., colloidal chemistry [18, 19], aerosol science (evolution of a system of solid or liquid particles suspended in a gas [6]), astrophysics [3], or hematology (red blood cell formation [16]). Assuming that each cluster is fully identified by its size (or volume), mean-field models have been developed and used to predict the time evolution of the size distribution function of the clusters. Various levels of description are also available within these models according to the range of the size parameter, which is either $\mathbb{N}\backslash\{0\}$ (discrete models) or $\mathbb{R}_+ = (0, +\infty)$ (continuous models). Among these models, the most widely used is the classical coagulation equation introduced by Smoluchowski (in its discrete version) to describe the aggregation of colloidal particles moving according to Brownian motion [18]. Since then, the Smoluchowski coagulation equation has been the subject of several physical and mathematical studies. In a different context, another coagulation model was proposed by Oort and van de Hulst thirty years later to describe the process of aggregation of protoplanetary bodies in astrophysics [15]. It was then written under a more tractable form by Safronov [17], to which we refer for a more detailed account on coagulation processes in astrophysics. Though the Smoluchowski and Oort–Hulst–Safronov (OHS) coagulation models were derived in different ways, it is natural to wonder whether there is some relationship between them. The main purpose of this work is actually to establish a connection between these two different coagulation equations. As a byproduct, we also obtain the existence of a solution to the OHS equation and study qualitative properties of this model.

Let us now state both models more precisely. If $f(u,t)$ denotes the density of clusters of size $u \in \mathbb{R}_+$ at time $t \geq 0$, then the classical continuous coagulation equation reads [6]

$$\partial_t f = Q_1(f)\,, \quad (u,t) \in \mathbb{R}_+^2\,, \tag{1.1}$$

*Received by the editors September 13, 2002; accepted for publication November 25, 2002; published electronically May 15, 2003. This research was supported by Polish KBN grant 2 PO3A 00717.
http://www.siam.org/journals/sima/34-6/41447.html

†Institute of Applied Mathematics and Mechanics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland (lachowic@mimuw.edu.pl, darekw@mimuw.edu.pl).

‡Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, Université Paul Sabatier–Toulouse 3, 118 route de Narbonne, F-31062 Toulouse cedex 4, France (laurenco@mip.ups-tlse.fr).

$$f(0) = f_0\,, \quad u \in \mathbb{R}_+\,, \tag{1.2}$$

where

$$Q_1(f)(u) = \frac{1}{2}\int_0^u a(u-v,v)\ f(u-v)\ f(v)\ dv - f(u)\ \int_0^\infty a(u,v)\ f(v)\ dv\,, \quad u \in \mathbb{R}_+\,, \tag{1.3}$$

and ∂_t denotes the partial derivative with respect to time. The reaction rate a in $Q_1(f)$ is a nonnegative function usually called the coagulation kernel and satisfies the following symmetry property:

$$0 \le a(u,v) = a(v,u)\,, \quad (u\,,v)\ \in \mathbb{R}_+^2\,. \tag{1.4}$$

The first term in $Q_1(f)$ is a gain term which describes the rate of formation of clusters of size u due to the merging of smaller clusters as a result of binary collisions. The second term in $Q_1(f)$ is a loss term accounting for the depletion of clusters of size u resulting from their coalescence with other clusters.

With the same notation, we may formulate the OHS model as follows [17]:

$$\partial_t f = Q_0(f)\,, \quad (u,t) \in \mathbb{R}_+^2\,, \tag{1.5}$$

along with the initial condition (1.2), where

$$Q_0(f)(u) = -\partial_u\left(f(u)\ \int_0^u v\ a(u,v)\ f(v)\ dv\right) - f(u)\ \int_u^\infty a(u,v)\ f(v)\ dv\,, \quad u \in \mathbb{R}_+\,. \tag{1.6}$$

Here ∂_u denotes the partial derivative with respect to u. Notice that if we omit the second term of $Q_0(f)$ in (1.5), the remaining part is a continuity equation which describes the increase of clusters of size u with velocity

$$\frac{du}{dt} = \int_0^u v\ a(u,v)\ f(v,t)\ dv$$

depending on the density of smaller clusters. Thus, in this model, the rate of formation of clusters of size u from smaller clusters does not depend on the sizes of the clusters involved in the coagulation event but on some averaged quantity, and this is a fundamental difference with (1.1). The second term of $Q_0(f)$ corresponds to the depletion of clusters of size u, which is here possible only as a result of their "sedimentation" on larger clusters. Another qualitative difference between both models is related to the speed with which an initial perturbation propagates. It is well known that the Smoluchowski equation (1.1) enjoys the property of infinite speed of propagation. (That is, if f_0 is compactly supported, the solution $f(t)$ to (1.1), (1.2) is not compactly supported for any positive time $t > 0$. We refer to [2] for a proof in the case of the discrete Smoluchowski equations.) This is in contrast with the OHS equation, where the propagation velocity is finite, as already observed in [7, 8] (see also section 5 below).

Nevertheless, our main goal is to show rigorously that it is possible to connect (1.1) to (1.5). Let us first mention that a relationship between these models has already

been observed by Dubovski [7] at a formal level. More precisely, Dubovski introduces a family of generalized discrete coagulation equations which includes the discrete Smoluchowski equations on the one hand and a discrete version of the OHS equation on the other hand. The connection is then completed by showing formally that the discrete version of the OHS equation leads to the OHS equation after a suitable rescaling, since it follows from [13] that (1.1) can be obtained as a limit of suitably rescaled discrete Smoluchowski equations. Our approach is completely different and directly connects (1.1) to (1.5) without using discrete models. We actually introduce an ε-dependent family of generalized coagulation equations for $\varepsilon \in (0\,,1]$. While $\varepsilon = 1$ corresponds to the Smoluchowski equation (1.1), letting $\varepsilon \to 0$ leads us to the OHS equation (1.5). Heuristically our approach is based on the following observation. Given a test function $\phi \in \mathcal{D}([0,+\infty))$, it follows from (1.3) and the Fubini theorem that

$$\int_0^\infty Q_1(f)\ \phi\ du = \int_0^\infty \int_0^v \left[\phi(v+w) - \phi(w) - \phi(v)\right] a(v,w) f(v) f(w)\ dwdv\,. \tag{1.7}$$

Similarly, (1.6) yields

$$\int_0^\infty Q_0(f)\ \phi\ du = \int_0^\infty \int_0^v \left[w\ \phi'(v) - \phi(w)\right]\ a(v,w)\ f(v)\ f(w)\ dwdv\,. \tag{1.8}$$

Observe now that for $v \gg w$, we have

$$\left[\phi(v+w) - \phi(w) - \phi(v)\right] \sim \left[w\ \phi'(v) - \phi(w)\right]\,, \tag{1.9}$$

so that we expect to recover (1.5) from (1.1) when the dominating reactions are the coalescence of clusters with very different sizes. Providing the rigorous justification of this observation is the first aim of our work and will be achieved with the help of the so-called generalized Boltzmann equations [1]. In order to see both equations in a common frame, we introduce the following generalized coagulation equation:

$$\partial_t f = Q_{GC}(f)\,, \quad (u,t) \in \mathbb{R}^2_+\,, \tag{1.10}$$

where

$$\begin{aligned} Q_{GC}(f)(u) = \frac{1}{2}\ &\int_0^\infty \int_0^\infty A(u;v,w)\ a(v,w)\ f(v)\ f(w)\ dwdv \\ &- f(u)\ \int_0^\infty a(u,v)\ f(v)\ dv\,. \end{aligned} \tag{1.11}$$

Here, A is the weighted probability that the collision of a cluster of size v and a cluster of size w generates a cluster of size u and is a nonnegative function satisfying

$$A(u;v,w) = A(u;w,v)\,, \quad (u,v,w) \in \mathbb{R}^3_+\,, \tag{1.12}$$

$$\int_0^\infty A(u;v,w)\ u\ du = v + w\,, \quad (v,w) \in \mathbb{R}^2_+\,. \tag{1.13}$$

It is worth mentioning here that many bilinear equations and systems in applied mathematics fit into this general structure after specifying appropriately A and a (cf. [11]). Observe also that the condition (1.13) ensures that the total volume is

preserved during the coagulation reactions. Indeed, we have

$$\int_0^\infty Q_{GC}(f)\ \phi\ du = \int_0^\infty \int_0^v \left\{ \left(\int_0^\infty A(u;v,w)\ \phi(u)\ du \right) - \phi(v) - \phi(w) \right\} a(v,w)\ f(v)\ f(w)\ dwdv \tag{1.14}$$

for any test function $\phi \in \mathcal{D}([0,+\infty))$. Setting now $\phi(u) = u$ in (1.14) and using (1.13), we obtain (at least formally)

$$\int_0^\infty Q_{GC}(f)\ u\ du = 0\,.$$

We thus formally deduce from (1.10) that, for $t > 0$,

$$\int_0^\infty u\ f(u,t)\ du = \int_0^\infty u\ f_0(u)\ du\,, \tag{1.15}$$

which is nothing but the conservation of the total mass throughout time evolution.

We are now in a position to introduce a family of generalized coagulation equations connecting the Smoluchowski and the OHS equations. For $\varepsilon \in (0\,,1)$ and $(v,w) \in \mathbb{R}_+^2$, we define

$$\begin{aligned} A_\varepsilon(u;v,w) = {} & \delta\left(u - \max\{v,w\} - \varepsilon\ \min\{v,w\}\right) \\ & + (1-\varepsilon)\ \delta(u - \min\{v,w\})\,, \end{aligned} \tag{1.16}$$

$$a_\varepsilon(v,w) = \frac{a(v,w)}{\varepsilon}\,, \tag{1.17}$$

where $\delta(u)$ is the Dirac mass. We next put A_ε instead of A and a_ε instead of a in (1.14) and set $Q_{GC} = Q_\varepsilon$. With this notation, (1.14) yields

$$\begin{aligned} & \int_0^\infty Q_\varepsilon(f)\ \phi\ du \\ & \quad = \int_0^\infty \int_0^v \left\{ \frac{\phi(v+\varepsilon w) - \phi(v)}{\varepsilon} - \phi(w) \right\} a(v,w)\ f(v)\ f(w)\ dwdv \end{aligned} \tag{1.18}$$

for any test function $\phi \in \mathcal{D}([0,+\infty))$. It is then straightforward to see that the choice $\varepsilon = 1$ in (1.18) yields (1.7), while letting $\varepsilon \to 0$ in (1.18) leads to (1.8). We have thus obtained the announced connection between (1.1) and (1.5). In fact, $Q_\varepsilon(f)$ may be written in a more explicit form, namely,

$$\begin{aligned} Q_\varepsilon(f)(u) = {} & \frac{1}{\varepsilon} \int_0^{u/(1+\varepsilon)} a(u-\varepsilon v, v)\ f(u-\varepsilon v)\ f(v)\ dv \\ & - f(u) \left[\left(\frac{1}{\varepsilon} - 1 \right) \int_0^u a(u,v)\ f(v)\ dv + \int_0^\infty a(u,v)\ f(v)\ dv \right] \end{aligned} \tag{1.19}$$

for $u \in \mathbb{R}_+$. We will actually prove the convergence of the solution f_ε to

$$\partial_t f_\varepsilon = Q_\varepsilon(f_\varepsilon)\,, \quad (u,t) \in \mathbb{R}_+^2\,, \tag{1.20}$$

with $f_\varepsilon(0) = f_0$ toward a solution to (1.5), (1.2) as $\varepsilon \to 0$.

Another issue we consider in this paper is the validity of the total mass conservation (1.15) which is derived formally from (1.14). It is well known that it is not always true for the classical Smoluchowski coagulation equation (1.1) and that the total mass may decrease after some time, a phenomenon known as gelation (see [5, 9, 10] and the references therein). The occurrence of gelation depends heavily on the growth of the coagulation kernel a, and the situation is quite clear for (1.1). We obtain here a similar result for the OHS equation by the arguments of [9] and investigate the connection between gelation and the propagation of the support for compactly supported initial data. We thereby extend results from [7, section 5] to a wider class of coagulation kernels.

We now describe the contents of this paper: The next section is devoted to a precise statement of our results, including convergence of solutions to (1.20) toward a solution to (1.5), occurrence of gelation, and behavior of the support for compactly supported initial data. The convergence proof is performed in section 3, and the occurrence of gelation is studied in section 4. Compactly supported initial data are the subject of the final section.

2. Main results. We first introduce the basic assumptions on the data a and f_0 that we shall use in what follows. Besides the nonnegativity and symmetry condition (1.4), we assume that the coagulation kernel a satisfies

$$a \in W^{1,\infty}_{\rm loc}([0,+\infty)^2) \quad \text{and} \quad \partial_u a(u,v) \ge -\alpha\,, \quad (u,v) \in \mathbb{R}^2_+\,, \tag{2.1}$$

for some constant $\alpha \ge 0$. We also need to prescribe the behavior of a for large values of u and v. A natural growth condition is to say that the rate of cluster interactions is limited by the product of their sizes or volumes, that is,

$$a(u,v) \le K\ (1+u)\ (1+v)\,, \quad (u,v) \in \mathbb{R}^2_+\,, \tag{2.2}$$

which includes most of the cases considered in the physical literature. However, under the sole growth condition (2.2), the existence of a solution is still an open problem even for (1.1), so that additional growth conditions have to be specified. Two different sets of growth assumptions will actually be used in what follows. Namely, in addition to (2.2), we will require that either a is *strictly subquadratic*, that is,

$$\omega_R(v) = \sup_{u\in[0,R]} \frac{a(u,v)}{v} \longrightarrow 0 \quad \text{as} \quad v \to +\infty \tag{2.3}$$

for each $R \ge 1$, or a is *subadditive*, that is, there is $K_1 > 0$ such that

$$a(u,v) \le K_1\ (1+u+v)\,, \quad (u,v) \in \mathbb{R}^2_+\,. \tag{2.4}$$

We turn next to the initial datum f_0 and notice that physically relevant requirements on f_0 are nonnegativity and finite total mass. We thus assume that the initial datum f_0 satisfies

$$f_0 \in X^+\,, \tag{2.5}$$

where X^+ denotes the positive cone of the Banach space

$$X = L^1(0,+\infty;(1+u)du)\,,$$

endowed with the norm

$$\|f\|_X = \int_0^\infty |f(u)|\ (1+u)\ du\,.$$

Notice that the last term in the norm $\|.\|_X$ corresponds to the total mass of clusters.

We now give the definition of a weak solution we use in this paper.

DEFINITION 2.1. *Let $T \in (0\,,+\infty]$ and $f_0 \in X^+$. A weak solution to* (1.5), (1.2) *on $[0,T)$ is a nonnegative function*

$$f \in \mathcal{C}([0\,,T); weak - L^1(\mathbb{R}_+)) \cap L^\infty(0,T;X^+), \tag{2.6}$$

which satisfies

$$\begin{aligned}\int_0^\infty f(u,t)\ \phi(u)\ du &= \int_0^\infty f_0(u)\ \phi(u)\ du \\ &+ \int_0^t \int_0^\infty \int_0^v [w\ \phi'(v) - \phi(w)]\ a(v,w)\ f(v,s)\ f(w,s)\ dwdvds\end{aligned} \tag{2.7}$$

for any $\phi \in W^{1,\infty}(\mathbb{R}_+)$ with compactly supported first derivative and $t \in [0,T)$.

Similarly, for $\varepsilon \in (0\,,1]$, a weak solution to (1.20), (1.2) *on $[0,T)$ is a nonnegative function*

$$f_\varepsilon \in \mathcal{C}([0\,,T); weak - L^1(\mathbb{R}_+)) \cap L^\infty(0,T;X^+)$$

satisfying

$$\begin{aligned}\int_0^\infty f_\varepsilon(u,t)\ \phi(u)\ du &= \int_0^\infty f_0(u)\ \phi(u)\ du \\ &+ \int_0^t \int_0^\infty \int_0^v \left\{ \frac{\phi(v+\varepsilon w) - \phi(v)}{\varepsilon} - \phi(w) \right\} a(v,w) f_\varepsilon(v,s) f_\varepsilon(w,s)\ dwdvds\end{aligned} \tag{2.8}$$

for any $\phi \in L^\infty(\mathbb{R}_+)$ and $t \in [0,T)$.

Our first result then reads as follows.

THEOREM 2.2. *Assume that the coagulation kernel a satisfies* (1.4), (2.1), *and either* (2.3) *or* (2.4). *Then, for $\varepsilon \in (0,1]$ and $f_0 \in X^+$, there exists at least one weak solution f_ε to* (1.20), (1.2) *on $[0,+\infty)$ which satisfies*

$$\int_0^\infty u\ f_\varepsilon(u,t)\ du \le \int_0^\infty u\ f_0(u)\ du\,, \quad t \ge 0\,, \tag{2.9}$$

if (2.3) *holds and*

$$\int_0^\infty u\ f_\varepsilon(u,t)\ du = \int_0^\infty u\ f_0(u)\ du\,, \quad t \ge 0\,, \tag{2.10}$$

if (2.4) *holds.*

Furthermore, there is a subsequence $\varepsilon_k \to 0$ and a weak solution f to (1.5), (1.2) *on $[0,+\infty)$ such that*

$$f_{\varepsilon_k} \longrightarrow f \quad in \quad \mathcal{C}([0\,,T]; weak - L^1(\mathbb{R}_+)) \tag{2.11}$$

for each $T > 0$ and f satisfies (2.9) *if* (2.3) *holds and* (2.10) *if* (2.4) *holds.*

Here and below, if B is a Banach space and $T \in (0,+\infty)$, $\mathcal{C}([0\,,T];weak-B)$ denotes the space of weakly continuous functions from $[0,T]$ in B.

For the classical Smoluchowski coagulation equation (1.1) ($\varepsilon = 1$), the existence results stated in Theorem 2.2 are already known; see [20, 13] and the references therein.

Remark 2.3. Under the assumption (2.4), the convergence (2.11) can actually be improved to

$$f_{\varepsilon_k} \longrightarrow f \quad \text{in} \quad \mathcal{C}([0\,,T];weak-X). \tag{2.12}$$

If f_0 is bounded, the solution constructed in Theorem 2.2 is also bounded, as we state now.

PROPOSITION 2.4. *Assume that the assumptions of Theorem* 2.2 *are fulfilled and also that* $f_0 \in L^\infty(\mathbb{R}_+)$. *Then the solution* f_ε *to* (1.20), (1.2) *for* $\varepsilon \in (0,1]$ *and the solution* f *to* (1.5), (1.2) *constructed in Theorem* 2.2 *satisfy*

$$\|f_\varepsilon(t)\|_{L^\infty}\,,\ \|f(t)\|_{L^\infty} \le \|f_0\|_{L^\infty}\ \exp\left(\alpha\ t\ \int_0^\infty u\ f_0(u)\ du\right) \tag{2.13}$$

for $t \ge 0$.

Next we turn to the occurrence of gelation and define the gelation time $T_{gel} \in [0,+\infty]$ of a solution f to the OHS equation (1.5), (1.2) by

$$T_{gel} = \inf\left\{t \ge 0\,,\quad \int_0^\infty u\ f(u,t)\ du < \int_0^\infty u\ f_0(u)\ du\right\}. \tag{2.14}$$

The occurrence of gelation then corresponds to $T_{gel} < +\infty$. Clearly, Theorem 2.2 ensures that there are mass-conserving solutions to the OHS equation (1.5) for subadditive coagulation kernels, the same result being true for the classical Smoluchowski coagulation equation. We now show that gelation takes place for (1.5) for the same class of coagulation kernels as for (1.1).

THEOREM 2.5. *Assume that a satisfies* (1.4), (2.2) *and*

$$a(u,v) \ge K_0\ (uv)^{\lambda/2}\,,\quad (u,v) \in \mathbb{R}_+^2, \tag{2.15}$$

for some $\lambda \in (1,2]$ *and* $K_0 > 0$. *Consider* $f_0 \in X^+$, $f_0 \not\equiv 0$, *and denote by* f *a weak solution to* (1.5), (1.2). *Then* $T_{gel} < +\infty$.

The proof of Theorem 2.5 is similar to that of [9, Theorem 1.1] for (1.1). It is worth mentioning that the method developed in [9] provides actually much more information on (1.1) than the mere occurrence of gelation (temporal decay estimates, behavior of higher moments near the gelation time, etc.). It is likely that similar results are valid for (1.5) with the same proofs, and we refer to [9] for a more detailed description of available results.

We finally consider compactly supported initial data. As already mentioned, a striking difference between (1.1) and (1.5) is that solutions to the latter enjoy the property of finite speed of propagation. More precisely, we have the following result.

THEOREM 2.6. *Assume that a satisfies* (1.4), (2.2) *and is continuous on* $[0,+\infty)^2$. *We consider* $f_0 \in X^+ \cap L^\infty(\mathbb{R}_+)$ *such that*

$$\text{supp}\ \ f_0 \subset [0,R_0] \tag{2.16}$$

for some $R_0 > 0$ *and denote by* f *a weak solution to* (1.5), (1.2) *such that* $f \in L^\infty((0,T)\times\mathbb{R}_+)$ *for each* $T \ge 0$. *Then there are* $T_\star \in (0,+\infty]$ *and* $R \in \mathcal{C}^1([0,T_\star))$ *such that*

(a) $R(0) = R_0$ *and* R *satisfies*

$$R'(t) = \int_0^{R(t)} u\ a(R(t),u)\ f(u,t)\ du\,, \quad t \in [0,T_\star)\,. \tag{2.17}$$

In addition, R *is a nondecreasing function on* $[0,T_\star)$ *and either* $T_\star = +\infty$ *or*

$$T_\star < +\infty \quad \text{and} \quad \lim_{t\to T_\star} R(t) = +\infty\,. \tag{2.18}$$

(b) supp $f(t) \subset [0,R(t)]$ *for* $t \in [0,T_\star)$.

In particular, when $f_0 \in L^\infty(\mathbb{R}_+)$, Theorem 2.6 applies to the solution f to (1.5), (1.2) constructed in Theorem 2.2 by Proposition 2.4. Next, some information on the time behavior of the total mass readily follows from Theorem 2.6 and is gathered below.

COROLLARY 2.7. *Under the assumptions and notation of Theorem* 2.6, *we have*

$$\int_0^\infty u\ f(u,t)\ du = \int_0^\infty u\ f_0(u)\ du \quad \text{for} \quad t \in [0,T_\star)\,, \tag{2.19}$$

and thus $T_\star \le T_{gel}$.

Since we already know that gelation takes place for coagulation kernels satisfying (2.15), we conclude that $T_\star < +\infty$ in that case, thereby extending [7, section 5]. Let us also mention that a natural conjecture is of course that $T_\star = T_{gel}$, but we have not been able to prove it.

3. Existence of weak solutions. The existence proof relies on the construction of an approximating sequence of solutions to some regularized problems combined with weak compactness arguments in $L^1(\mathbb{R}_+)$. Such an approach was introduced in [20] for the classical coagulation equation (1.1) and further developed in [13]. We will adapt arguments from both papers to prove Theorem 2.2.

We first derive the following a priori estimates.

LEMMA 3.1. *Let* $f_0 \in X^+$ *and* $\varepsilon \in (0,1]$. *Assume that there is a nonnegative and convex piecewise* $\mathcal{C}^2$*-function* $\Phi \in \mathcal{C}^1([0\,,+\infty))$ *such that* $\Phi(0) = 0$, $\Phi'(0) = 1$, Φ' *is concave, and*

$$C_0 = \int_0^\infty \Phi(f_0)(u)\ du < \infty\,. \tag{3.1}$$

Let $T > 0$ *and* $f_\varepsilon \in \mathcal{C}^1([0,T);L^1(\mathbb{R}_+))$ *be a weak solution to* (1.20), (1.2) *on* $[0,T)$. *There is a positive constant* C_1 *depending only on* C_0, T, α *in* (2.1) *and* $\|f_\varepsilon\|_{L^\infty(0,T;X)}$ *such that*

$$\sup_{t\in[0\,,T)} \int_0^\infty \Phi(f_\varepsilon(u,t))\ du \le C_1\,. \tag{3.2}$$

Proof. We first recall (cf. [12, Lemma A.1]) that the properties of Φ imply

$$x\ \Phi'(x) \le 2\ \Phi(x) \quad \text{for} \quad x \ge 0\,. \tag{3.3}$$

Next we introduce

$$\Phi_R(x) = \begin{cases} \Phi(x) & \text{if} \quad x \in [0\,,R]\,, \\ \Phi'(R)(x-R) + \Phi(R) & \text{if} \quad x \in [R\,,+\infty) \end{cases} \tag{3.4}$$

for $R \ge 2$ and notice that Φ_R has a bounded first derivative, $\Phi_R \le \Phi$, and also Φ_R enjoys the same properties as Φ and in particular (3.3).

We multiply (1.20) by $\Phi_R'(f_\varepsilon)$, integrate on $\mathbb{R}_+$, and use (1.18) with $\phi = \Phi_R'(f_\varepsilon)$ to obtain

$$\begin{aligned}
&\frac{d}{dt}\int_0^\infty \Phi_R(f_\varepsilon)\ du \\
&\le \int_0^\infty \int_0^v \left\{ \frac{\Phi_R'(f_\varepsilon(v+\varepsilon w)) - \Phi_R'(f_\varepsilon(v))}{\varepsilon} \right\}\ a(v,w)\ f_\varepsilon(v)\ f_\varepsilon(w)\ dwdv\,,
\end{aligned}$$

since the monotonicity of Φ_R ensures that the last term of the right-hand side of (1.18) gives a nonpositive contribution. The convexity of Φ_R next entails that for $x,\ y \ge 0$,

$$x\ (\Phi_R'(y) - \Phi_R'(x)) \le y\ \Phi_R'(y) + \Phi_R(x) - \Phi_R(y) - x\ \Phi_R'(x) = \Psi_R(y) - \Psi_R(x)\,,$$

where $\Psi_R(x) = x\ \Phi_R'(x) - \Phi_R(x)$, $x \ge 0$, is a nonnegative function since Φ_R is convex with $\Phi_R(0) = 0$. Consequently,

$$\begin{aligned}
&\frac{d}{dt}\int_0^\infty \Phi_R(f_\varepsilon)\ du \\
&\le \frac{1}{\varepsilon}\ \int_0^\infty \int_0^v (\Psi_R(f_\varepsilon(v+\varepsilon w)) - \Psi_R(f_\varepsilon(v)))\ a(v,w)\ f_\varepsilon(w)\ dwdv \\
&\le \frac{1}{\varepsilon}\ \int_0^\infty f_\varepsilon(w) \int_w^\infty (\Psi_R(f_\varepsilon(v+\varepsilon w)) - \Psi_R(f^\varepsilon(v)))\ a(v,w)\ dvdw \\
&\le \frac{1}{\varepsilon}\ \int_0^\infty \int_{(1+\varepsilon)w}^\infty a(v-\varepsilon w, w)\ \Psi_R(f_\varepsilon(v))\ f_\varepsilon(w)\ dvdw \\
&\quad - \frac{1}{\varepsilon}\ \int_0^\infty \int_w^\infty a(v,w)\ \Psi_R(f_\varepsilon(v))\ f_\varepsilon(w)\ dvdw \\
&\le \int_0^\infty \int_w^\infty \left\{ \frac{a(v-\varepsilon w, w) - a(v,w)}{\varepsilon} \right\}\ \Psi_R(f_\varepsilon(v))\ f_\varepsilon(w)\ dvdw\,.
\end{aligned}$$

Next we use (2.1) and (3.3) to deduce that

$$\begin{aligned}
\frac{d}{dt}\int_0^\infty \Phi_R(f_\varepsilon)\ du &\le \alpha \int_0^\infty \int_0^\infty w\ \Psi_R(f_\varepsilon(v))\ f_\varepsilon(w)\ dwdv \\
&\le \alpha\ \|f_\varepsilon\|_{L^\infty(0,T;X)}\ \int_0^\infty \Phi_R(f_\varepsilon(v))\ dv\,.
\end{aligned}$$

We now apply the Gronwall lemma and let $R \to +\infty$ with the help of (3.1) to complete the proof. □

We next prove that for subadditive coagulation kernels, moments propagate throughout time evolution.

LEMMA 3.2. *Let $f_0 \in X^+$ and $\varepsilon \in (0,1]$. Assume that there is a nonnegative and convex piecewise $\mathcal{C}^2$-function $\varphi \in \mathcal{C}^1([0\,,+\infty))$ such that $\varphi(0) = 0$, $\varphi'(0) = 1$, φ' is concave, and*

$$C_2 = \int_0^\infty \varphi(u)\ f_0(u)\ du < \infty\,. \tag{3.5}$$

Assume further that the coagulation kernel a satisfies (2.4), and let $T > 0$ and $f_\varepsilon \in \mathcal{C}^1([0,T); L^1(\mathbb{R}_+))$ be a weak solution to (1.20), (1.2) on $[0,T)$. There is a positive

constant C_3 depending only on C_2, T, K_1 in (2.4), $\|\varphi\|_{L^\infty(0,2)}$, and $\|f_\varepsilon\|_{L^\infty(0,T;X)}$ such that

$$\sup_{t\in[0\,,T)} \int_0^\infty \varphi(u)\ f_\varepsilon(u,t)\ du \le C_3\,. \tag{3.6}$$

Proof. We first recall (cf. [12, Lemma A.2]) that the properties of φ imply

$$(x+y)\ (\varphi(x+y)-\varphi(x)-\varphi(y)) \le 2\ (x\ \varphi(y)+y\ \varphi(x))\ , \quad x,y\ge 0\,. \tag{3.7}$$

Owing to the convexity of φ, we have

$$\begin{aligned}\varphi(v+\varepsilon w)-\varphi(v) &\le \varepsilon\ \varphi(v+w)+(1-\varepsilon)\ \varphi(v)-\varphi(v)\\ &\le \varepsilon\ (\varphi(v+w)-\varphi(v))\end{aligned}$$

for $(v,w)\in\mathbb{R}^2_+$, and it follows from (1.18) with $\phi=\varphi$, (2.4), and (3.7) that

$$\begin{aligned}&\frac{d}{dt}\int_0^\infty \varphi(u)\ f_\varepsilon\ du\\ &\le \int_0^\infty\int_0^v (\varphi(v+w)-\varphi(v)-\varphi(w))\ a(v,w)\ f_\varepsilon(v)\ f_\varepsilon(w)\ dwdv\\ &\le 3\ K_1\ \|\varphi\|_{L^\infty(0,2)}\ \int_0^1\int_0^v f_\varepsilon(v)\ f_\varepsilon(w)\ dwdv\\ &\quad +\ 2\ K_1\ \int_1^\infty\int_0^v (v+w)\ (\varphi(v+w)-\varphi(v)-\varphi(w))\ f_\varepsilon(v)\ f_\varepsilon(w)\ dwdv\\ &\le C_3\ \|f_\varepsilon\|^2_{L^\infty(0,T;X)}+4\ K_1\ \int_0^\infty\int_0^\infty (v\ \varphi(w)+w\ \varphi(v))\ f_\varepsilon(v)\ f_\varepsilon(w)\ dwdv\\ &\le C_3+8\ K_1\ \|f_\varepsilon\|_{L^\infty(0,T;X)}\ \int_0^\infty \varphi(v)\ f_\varepsilon(v)\ dv\,.\end{aligned}$$

Recalling (3.5), we conclude by the Gronwall lemma that (3.6) holds true. The above computation is actually mostly formal as f_ε might not be integrable against the weight φ. A rigorous justification can be performed by approximating φ by Lipschitz continuous functions, as in the proof of Lemma 3.1. □

Before turning to the proof of Theorem 2.2, let us recall the following lemma.

LEMMA 3.3. *Let $p>0$, $(\psi_n)_{n\ge1}$, $\psi\in L^\infty((0\,,p)\times(0\,,p))$, and $(g_n)_{n\ge1}$, $g\in L^1(0\,,p)$. Suppose that*

$$\sup_{n\ge1}\|\psi_n\|_{L^\infty((0\,,p)\times(0\,,p))}<+\infty$$

and

$$\begin{aligned}\psi_n &\longrightarrow \psi \quad \text{a.e. in} \quad (0\,,p)\times(0\,,p)\,,\\ g_n &\rightharpoonup g \quad \text{weakly in } L^1(0\,,p)\,.\end{aligned}$$

Then

$$\lim_{n\to+\infty}\int_0^p\int_0^p \psi_n(x,y)\ g_n(x)\ g_n(y)\ dxdy=\int_0^p\int_0^p \psi(x,y)\ g(x)\ g(y)\ dxdy\,.$$

The proof of Lemma 3.3 relies on the Dunford–Pettis and Egorov theorems and is contained implicitly in [20, Lemma 4.1], to which we refer.

Proof of Theorem 2.2. Throughout the proof, we denote by C any positive constant which depends only on α in (2.1), K in (2.2), and f_0. The dependence of C upon additional parameters will be indicated explicitly.

We fix $\varepsilon \in (0\,,1]$ and $\varrho \ge 1$. We first consider a regularized problem obtained by substituting a in (1.19) for a_ϱ defined by

$$a_\varrho(v,w) = \min\{a(v,w),\varrho\}\,, \quad (v,w) \in \mathbb{R}_+^2\,, \tag{3.8}$$

denoting by $Q_{\varepsilon,\varrho}$ the coagulation operator thus obtained. Introducing

$$Q^1_{\varepsilon,\varrho}(f)(u) = \frac{1}{\varepsilon} \int_0^{u/(1+\varepsilon)} a_\varrho(u-\varepsilon v, v)\ f(u-\varepsilon v)\ f(v)\ dv\,, \quad u \in \mathbb{R}_+\,,$$

and $Q^2_{\varepsilon,\varrho}(f) = Q_{\varepsilon,\varrho}(f) - Q^1_{\varepsilon,\varrho}(f)$, it is easy to check that the boundedness of a_ϱ ensures that $Q^i_{\varepsilon,\varrho}$, $i = 1,2$, are locally Lipschitz continuous functions from $L^1(\mathbb{R}_+)$ in $L^1(\mathbb{R}_+)$. Using the Banach fixed point theorem, one proves that there exist $T_{\varepsilon,\varrho} \in (0,+\infty]$ and a unique solution $f_{\varepsilon,\varrho} \in \mathcal{C}^1([0,T_{\varepsilon,\varrho}); L^1(\mathbb{R}_+))$ to

$$\partial_t f_{\varepsilon,\varrho} = \left(Q^1_{\varepsilon,\varrho}(f_{\varepsilon,\varrho})\right)_+ + Q^2_{\varepsilon,\varrho}(f_{\varepsilon,\varrho})\,, \quad (u,t) \in \mathbb{R}_+ \times [0,T_{\varepsilon,\varrho})\,, \tag{3.9}$$

with $f_{\varepsilon,\varrho}(0) = f_0$. Here and below, we denote by $x_+ = \max\{x,0\}$ the positive part of the real number x. Since f_0 and the first term of the right-hand side of (3.9) are nonnegative, we deduce from (3.9) that

$$f_{\varepsilon,\varrho}(t) \ge 0 \quad \text{a.e. in} \quad \mathbb{R}_+$$

for every $t \in [0,T_{\varepsilon,\varrho})$. Consequently, $\left(Q^1_{\varepsilon,\varrho}(f_{\varepsilon,\varrho})\right)_+ = Q^1_{\varepsilon,\varrho}(f_{\varepsilon,\varrho})$ and $f_{\varepsilon,\varrho}$ solves

$$\partial_t f_{\varepsilon,\varrho} = Q_{\varepsilon,\varrho}(f_{\varepsilon,\varrho})\,, \quad (u,t) \in \mathbb{R}_+ \times [0,T_{\varepsilon,\varrho})\,, \tag{3.10}$$

with $f_{\varepsilon,\varrho}(0) = f_0$. In addition, it follows at once from (1.18) with $\phi(v) = 1$ and (3.10) that

$$\int_0^\infty f_{\varepsilon,\varrho}(u,t)\ du \le \int_0^\infty f_0(u)\ du\,, \quad t \in [0\,,T_{\varepsilon,\varrho})\,, \tag{3.11}$$

whence $T_{\varepsilon,\varrho} = +\infty$. Next, let $t > 0$ and $R \ge 1$. It follows from (3.10) that

$$\begin{aligned}\int_0^\infty \min\{u,R\}\ f_{\varepsilon,\varrho}(u,t)\ du = &\int_0^\infty \min\{u,R\}\ f_0(u)\ du \\ &+ \int_0^t \int_0^\infty Q_{\varepsilon,\varrho}(f_{\varepsilon,\varrho}(u,s))\ \min\{u,R\}\ du ds\,.\end{aligned} \tag{3.12}$$

Since

$$\frac{\min\{v+\varepsilon w,R\} - \min\{v,R\}}{\varepsilon} - \min\{w,R\} \le 0\,, \quad 0 \le w \le v\,,$$

we infer from (1.18) with $\phi(v) = \min\{v,R\}$, $v \in \mathbb{R}_+$, that the last term on the right-hand side of (3.12) is nonpositive, and we conclude that

$$\int_0^\infty \min\{u,R\}\ f_{\varepsilon,\varrho}(u,t)\ du \le \int_0^\infty \min\{u,R\}\ f_0(u)\ du \le \int_0^\infty u\ f_0(u)\ du\,.$$

Consequently, $f_{\varepsilon,\varrho} \in L^\infty(0,+\infty;X)$ by the Fatou lemma, which, together with the boundedness of a_ϱ, entails that

$$\lim_{R\to+\infty} \int_0^t \int_0^\infty Q_{\varepsilon,\varrho}(f_{\varepsilon,\varrho}(u,s))\ \min\{u,R\}\ du ds = 0\,.$$

We may then let $R \to +\infty$ in (3.12) and obtain

$$\int_0^\infty u\ f_{\varepsilon,\varrho}(u,t)\ du = \int_0^\infty u\ f_0(u)\ du\,, \quad t \ge 0\,. \tag{3.13}$$

Gathering (3.11) and (3.13), we thus have shown that

$$\sup_{t\ge 0}\ \|f_{\varepsilon,\varrho}(t)\|_X \le \|f_0\|_X\,, \quad \varrho \ge 1\,. \tag{3.14}$$

Step 1. Compactness in $\mathcal{C}([0,T];weak - L^1(\mathbb{R}_+))$.

In this part of the proof, we require only the coagulation kernel a to fulfill (1.4), (2.1), and (2.2). Recall that, since $f_0 \in L^1(\mathbb{R}_+)$, it follows from a refined version of the de la Vallée–Poussin theorem [4, 14] that there exists a function Φ fulfilling the assumptions of Lemma 3.1 and such that

$$\frac{\Phi(x)}{x} \to +\infty \quad \text{as } x \to +\infty \quad \text{and} \quad \int_0^\infty \Phi(f_0(u))\ du < +\infty\,. \tag{3.15}$$

Also, we infer from (2.1) that

$$\partial_u a_\varrho(u,v) = \left(\frac{1 - \ \text{sign}\ (a(u,v)-\varrho)}{2}\right)\ \partial_u a(u,v) \ge -\alpha\,, \quad (u,v) \in \mathbb{R}_+^2\,. \tag{3.16}$$

Owing to (3.14), (3.15), and (3.16), we are in a position to apply Lemma 3.1 and conclude that for each $T > 0$,

$$\sup_{t\in[0,T]} \int_0^\infty \Phi(f_{\varepsilon,\varrho}(u,t))\ du \le C(T)\,, \tag{3.17}$$

uniformly with respect to $\varepsilon \in (0,1]$ and $\varrho \ge 1$. Thanks to (3.14), the superlinearity (3.15) of Φ, and (3.17), we infer from the Dunford–Pettis theorem that for each $T > 0$, there is a weakly compact subset $\mathcal{K}_T$ of $L^1(\mathbb{R}_+)$ such that

$$f_{\varepsilon,\varrho}(t) \in \mathcal{K}_T\,, \quad (t,\varepsilon,\varrho) \in [0,T] \times (0,1] \times [1,+\infty)\,. \tag{3.18}$$

We next show the time equicontinuity of the sequence $(f_{\varepsilon,\varrho})$ in the weak topology of $L^1(\mathbb{R}_+)$. We first consider $\phi \in \mathcal{D}(\mathbb{R}_+)$ and let $r_0 > 0$ be such that supp $\phi \subset [0,r_0]$. For $h > 0$ and $t \in [0,T-h]$, it follows from (1.18), (2.2), and (3.14) that

$$\begin{aligned}
&\left| \int (f_{\varepsilon,\varrho}(t+h) - f_{\varepsilon,\varrho}(t))\ \phi\ du \right| \\
&\le \int_t^{t+h} \int_0^{r_0} \int_0^v a(v,w)\ \left| \frac{\phi(v+\varepsilon\ w) - \phi(v)}{\varepsilon} \right|\ f_{\varepsilon,\varrho}(v)\ f_{\varepsilon,\varrho}(w)\ dw dv ds \\
&\quad + \int_t^{t+h} \int_0^\infty \int_0^\infty a(v,w)\ |\phi(w)|\ f_{\varepsilon,\varrho}(v)\ f_{\varepsilon,\varrho}(w)\ dv dw ds
\end{aligned}$$

$$
\begin{aligned}
&\le \|\partial_v \phi\|_{L^\infty} \int_t^{t+h} \int_0^{r_0} \int_0^v a(v,w)\ w\ f_{\varepsilon,\varrho}(v)\ f_{\varepsilon,\varrho}(w)\ dwdvds \\
&\quad + \|\phi\|_{L^\infty} \int_t^{t+h} \int_0^{\infty} \int_0^{\infty} a(v,w)\ f_{\varepsilon,\varrho}(v)\ f_{\varepsilon,\varrho}(w)\ dvdwds \\
&\le K\ (1+r_0)\ \|\phi\|_{W^{1,\infty}} \int_t^{t+h} \int_0^{\infty} \int_0^{\infty} (1+v)\ (1+w)\ f_{\varepsilon,\varrho}(v)\ f_{\varepsilon,\varrho}(w)\ dvdwds \\
&\le K\ (1+r_0)\ \|\phi\|_{W^{1,\infty}}\ \|f_{\varepsilon,\varrho}\|^2_{L^\infty(0,T;X)}\ |h| \\
&\le C(\phi)\ |h|\,.
\end{aligned}
$$

This proves the time equicontinuity for test functions ϕ in $\mathcal{D}(\mathbb{R}_+)$. For a general function in $L^\infty(\mathbb{R}_+)$, we use the following approximation argument: for a given $\phi \in L^\infty(0,+\infty)$, there is a sequence of functions (ϕ_k) in $\mathcal{D}(\mathbb{R}_+)$ such that $\|\phi_k\|_{L^\infty} \le 2\ \|\phi\|_{L^\infty}$ and

$$
\phi_k \longrightarrow \phi \quad \text{a.e. in} \quad \mathbb{R}_+.
$$

It then follows from the Egorov theorem that for each $R > 0$ and $\delta \in (0,1)$, there is a measurable subset $Z_{\delta,R}$ of $(0,R)$ with $|Z_{\delta,R}| \le \delta$ and such that

$$
\lim_{k\to+\infty} \sup_{v\in(0,R)\setminus Z_{\delta,R}} |\phi_k(v) - \phi(v)| = 0.
$$

Now, we have

$$
\begin{aligned}
&\left| \int_0^{\infty} (f_{\varepsilon,\varrho}(t+h) - f_{\varepsilon,\varrho}(t))\ \phi\ du \right| \\
&\le \left| \int_0^{\infty} (f_{\varepsilon,\varrho}(t+h) - f_{\varepsilon,\varrho}(t))\ \phi_k\ du \right| + \left| \int_0^{\infty} (f_{\varepsilon,\varrho}(t+h) - f_{\varepsilon,\varrho}(t))\ (\phi - \phi_k)\ du \right| \\
&\le C(\phi_k)\ |h| + \int_{(0,R)\setminus Z_{\delta,R}} (f_{\varepsilon,\varrho}(t+h) + f_{\varepsilon,\varrho}(t))\ |\phi_k - \phi|\ du \\
&\quad + 3\ \|\phi\|_{L^\infty} \sup_{t\in[0,T]} \int_R^{\infty} f_{\varepsilon,\varrho}(t)\ du + 3\ \|\phi\|_{L^\infty} \sup_{t\in[0,T]} \int_{Z_{\delta,R}} f_{\varepsilon,\varrho}(t)\ du \\
&\le C(\phi_k)\ |h| + 2\ \|f_{\varepsilon,\varrho}\|_{L^\infty(0,T;X)} \sup_{v\in(0,R)\setminus Z_{\delta,R}} |\phi_k(v) - \phi(v)| \\
&\quad + 3\ \|\phi\|_{L^\infty} \sup_{t\in[0,T]} \int_R^{\infty} f_{\varepsilon,\varrho}(t)\ du + 3\ \|\phi\|_{L^\infty} \sup_{t\in[0,T]} \int_{Z_{\delta,R}} f_{\varepsilon,\varrho}(t)\ du\,.
\end{aligned}
$$

We first let $h \to 0$ and use (3.14) to obtain

$$
\begin{aligned}
\limsup_{h\to 0} \left| \int_0^{\infty} (f_{\varepsilon,\varrho}(t+h) - f_{\varepsilon,\varrho}(t))\ \phi\ du \right| &\le 2\ \|f_0\|_X \sup_{v\in(0,R)\setminus Z_{\delta,R}} |\phi_k(v) - \phi(v)| \\
&\quad + \frac{3}{R}\ \|\phi\|_{L^\infty}\ \|f_0\|_X \\
&\quad + 3\ \|\phi\|_{L^\infty} \sup_{t\in[0,T]} \int_{Z_{\delta,R}} f_{\varepsilon,\varrho}(t)\ du
\end{aligned}
$$

for every k, R, and δ. We then let $k \to +\infty$, and next $\delta \to 0$ and $R \to +\infty$, to conclude that

$$
\lim_{h\to 0} \left| \int_0^{\infty} (f_{\varepsilon,\varrho}(t+h) - f_{\varepsilon,\varrho}(t))\ \phi\ du \right| = 0\,,
$$

this limit being uniform with respect to $\varepsilon \in (0\,,1]$, $\varrho > 1$, and $t \in [0\,,T]$. Notice that in order to pass to the limit as $\delta \to 0$ in the last term of the above inequality, we used (3.18) and the Dunford–Pettis theorem.

We are now in a position to apply a variant of the Ascoli theorem (see, e.g., [21, Theorem 1.3.2]) to deduce that $(f_{\varepsilon,\varrho})$ lies in a relatively compact subset of $\mathcal{C}([0\,,T];weak-L^1(\mathbb{R}_+))$. There are thus sequences $\varrho \to +\infty$ and $\varepsilon \to 0$ (not relabeled) and nonnegative functions f_ε, $f \in \mathcal{C}([0\,,T];weak-L^1(\mathbb{R}_+))$ such that

$$f_{\varepsilon,\varrho} \longrightarrow f_\varepsilon \quad \text{in} \quad \mathcal{C}([0\,,T];weak-L^1(\mathbb{R}_+)) \tag{3.19}$$

for each $T > 0$ and $\varepsilon \in (0,1]$, and

$$f_\varepsilon \longrightarrow f \quad \text{in} \quad \mathcal{C}([0\,,T];weak-L^1(\mathbb{R}_+)) \tag{3.20}$$

for each $T > 0$. As a first consequence of (3.19) and (3.20), it follows from (3.14) and the weak lower semicontinuity of $\|.\|_X$ that

$$\sup_{t\ge 0} \|f_\varepsilon(t)\|_X \le \|f_0\|_X \quad \text{and} \quad \sup_{t\ge 0} \|f(t)\|_X \le \|f_0\|_X \tag{3.21}$$

for $\varepsilon \in (0\,,1]$.

We next identify the equations satisfied by f_ε for $\varepsilon \in (0\,,1]$ and f. We recall that for $\varepsilon \in (0\,,1]$ and $\varrho > 1$, the function $(f_{\varepsilon,\varrho})$ satisfies the weak form (2.8) of (3.10), that is,

$$\int_0^\infty f_{\varepsilon,\varrho}(u,t)\ \phi(u)\ du = \int f_0(u)\ \phi(u)\ du + \mathcal{R}^1_{\varepsilon,\varrho}(\phi,t) - \mathcal{R}^2_{\varepsilon,\varrho}(\phi,t) \tag{3.22}$$

for every $\phi \in \mathcal{D}([0\,,+\infty))$ and $t \ge 0$, where

$$\mathcal{R}^1_{\varepsilon,\varrho}(\phi,t) = \int_0^t \int_0^\infty \int_0^v \left(\frac{\phi(v+\varepsilon w) - \phi(v)}{\varepsilon} \right)\ a_\varrho(v,w)\ f_{\varepsilon,\varrho}(v,s)\ f_{\varepsilon,\varrho}(w,s)\ dwdvds\,,$$

$$\mathcal{R}^2_{\varepsilon,\varrho}(\phi,t) = \int_0^t \int_0^\infty \int_0^v \phi(w)\ a_\varrho(v,w)\ f_{\varepsilon,\varrho}(v,s)\ f_{\varepsilon,\varrho}(w,s)\ dwdvds\,.$$

We now fix $\phi \in \mathcal{D}([0\,,+\infty))$ and $t > 0$ in the above formulae and pass to the limit in (3.22) first as $\varrho \to +\infty$ and then as $\varepsilon \to 0$. Let R_0 be such that $\operatorname{supp}\phi \subset [0, R_0]$. On the one hand, notice that (3.19) and (3.20) readily imply that

$$\lim_{\varrho\to+\infty} \int_0^\infty f_{\varepsilon,\varrho}(u,t)\ \phi(u)\ du = \int_0^\infty f_\varepsilon(u,t)\ \phi(u)\ du \tag{3.23}$$

for $\varepsilon \in (0,1]$ and

$$\lim_{\varepsilon\to 0} \int_0^\infty f_\varepsilon(u,t)\ \phi(u)\ du = \int_0^\infty f(u,t)\ \phi(u)\ du\,. \tag{3.24}$$

On the other hand,

$$\phi(v+\varepsilon w) - \phi(v) = 0 \quad \text{if} \quad (v,w) \notin (0,R_0)^2 \quad \text{and} \quad 0 \le w \le v\,,$$

so that $\mathcal{R}^1_{\varepsilon,\varrho}(\phi,t)$ reduces to an integral over $(0,t)\times(0,R_0)^2$. Owing to (2.2), (3.19), and the definition of a_ϱ, we are in a position to apply Lemma 3.3 and conclude that

$$\lim_{\varrho\to+\infty} \mathcal{R}^1_{\varepsilon,\varrho}(\phi,t) = \mathcal{R}^1_\varepsilon(\phi,t)\,, \tag{3.25}$$

where

$$\mathcal{R}^1_\varepsilon(\phi,t) = \int_0^t \int_0^\infty \int_0^v \left(\frac{\phi(v+\varepsilon w)-\phi(v)}{\varepsilon} \right) \, a(v,w) \, f_\varepsilon(v,s) \, f_\varepsilon(w,s) \, dwdvds \, .$$

Similarly, we deduce from (2.2), (3.20), and Lemma 3.3 that

$$\lim_{\varepsilon\to 0} \mathcal{R}^1_\varepsilon(\phi,t) = \int_0^t \int_0^\infty \int_0^v w \, \phi'(v) \, a(v,w) \, f(v,s) \, f(w,s) \, dwdvds \, . \tag{3.26}$$

It remains to pass to the limit in $\mathcal{R}^2_{\varepsilon,\varrho}(\phi,t)$. Observe that this term involves values of $f_{\varepsilon,\varrho}(v,s)$ for large values of v so that more precise information on the behavior of a for large values of v is needed. We thus now split the proof according to the assumed growth of a.

Step 2. Convergence for strictly subquadratic kernels.

In this step, we complete the proof of Theorem 2.2 when the coagulation kernel a satisfies (2.3) in addition to (2.1) and (2.2). On the one hand, we observe that, given $R > 1$, it follows as above from (2.2), (3.19), and Lemma 3.3 that

$$\begin{aligned} &\lim_{\varrho\to+\infty} \int_0^t \int_0^R \int_0^v \phi(w) \, a_\varrho(v,w) \, f_{\varepsilon,\varrho}(v,s) \, f_{\varepsilon,\varrho}(w,s) \, dwdvds \\ &= \int_0^t \int_0^R \int_0^v \phi(w) \, a(v,w) \, f_\varepsilon(v,s) \, f_\varepsilon(w,s) \, dwdvds \, . \end{aligned}$$

On the other hand, we infer from (2.3) and (3.14) that

$$\begin{aligned} &\int_0^t \int_R^\infty \int_0^v \phi(w) \, a_\varrho(v,w) \, f_{\varepsilon,\varrho}(v,s) \, f_{\varepsilon,\varrho}(w,s) \, dwdvds \\ &\le \|\phi\|_{L^\infty} \int_0^t \int_R^\infty \int_0^{R_0} a(w,v) \, f_{\varepsilon,\varrho}(v,s) \, f_{\varepsilon,\varrho}(w,s) \, dwdvds \\ &\le \|\phi\|_{L^\infty} \int_0^t \int_R^\infty \int_0^{R_0} v \, \omega_{R_0}(v) \, f_{\varepsilon,\varrho}(v,s) \, f_{\varepsilon,\varrho}(w,s) \, dwdvds \\ &\le C(\phi) \, \|\omega_{R_0}\|_{L^\infty(R,+\infty)} \, , \end{aligned}$$

whence

$$\lim_{R\to+\infty} \int_0^t \int_R^\infty \int_0^v \phi(w) \, a_\varrho(v,w) \, f_{\varepsilon,\varrho}(v,s) \, f_{\varepsilon,\varrho}(w,s) \, dwdvds = 0$$

by (2.3), uniformly with respect to $\varepsilon \in (0,1]$ and $\varrho \ge 1$. Similarly, we deduce from (2.3) and (3.21) that

$$\lim_{R\to+\infty} \int_0^t \int_R^\infty \int_0^v \phi(w) \, a(v,w) \, f_\varepsilon(v,s) \, f_\varepsilon(w,s) \, dwdvds = 0 \, .$$

The above analysis then yields that

$$\lim_{\varrho\to+\infty} \mathcal{R}^2_{\varepsilon,\varrho}(\phi,t) = \mathcal{R}^2_\varepsilon(\phi,t) \, , \tag{3.27}$$

where

$$\mathcal{R}^2_\varepsilon(\phi,t) = \int_0^t \int_0^\infty \int_0^v \phi(w) \, a(v,w) \, f_\varepsilon(v,s) \, f_\varepsilon(w,s) \, dwdvds \, .$$

Proceeding analogously, we next obtain that

$$\lim_{\varepsilon\to 0} \mathcal{R}^2_\varepsilon(\phi,t) = \int_0^t \int_0^\infty \int_0^v \phi(w)\ a(v,w)\ f(v,s)\ f(w,s)\ dwdvds\,. \tag{3.28}$$

Gathering (3.23), (3.25), and (3.27) ensures that f_ε satisfies (2.8) for any $\phi \in \mathcal{D}([0\,,+\infty))$, and (3.21) together with classical approximation arguments entails that f_ε actually satisfies (2.8) for any $\phi \in L^\infty(0\,,+\infty)$. Also, (3.13) and (3.19) warrant that (2.9) holds true. Next, gathering (3.24), (3.26), and (3.28) implies that f satisfies (2.7) for any $\phi \in \mathcal{D}([0\,,+\infty))$. Using again classical approximation arguments along with (3.21) then yields that f satisfies (2.7) for any $\phi \in W^{1,\infty}(0\,,+\infty)$ with compactly supported first derivative. Finally, the inequality (2.9) for f_ε and (3.20) yield (2.9) for f. Recalling the convergence (3.20), we have proved Theorem 2.2 for strictly subquadratic coagulation kernels a.

Step 3. Convergence for subadditive kernels.

This final step is devoted to the proof of Theorem 2.2 for a subadditive coagulation kernel a satisfying (2.4) besides (2.1) and (2.2). As in the previous step, the main issue is to pass to the limit in $\mathcal{R}^2_{\varepsilon,\varrho}(\phi,t)$. For that purpose, we will employ Lemma 3.2 and use once more a refined version of the de la Vallée–Poussin theorem [4, 14] to find a function φ enjoying the requirements of Lemma 3.2 and

$$\frac{\varphi(x)}{x} \to +\infty \quad \text{as } x \to +\infty \quad \text{and} \quad \int_0^\infty \varphi(u)\ f_0(u)\ du < +\infty\,. \tag{3.29}$$

Since a_ϱ fulfills (2.4) with the same constant K_1 as a, we infer from Lemma 3.2 that

$$\sup_{t\in[0,T]} \int_0^\infty \varphi(u)\ f_{\varepsilon,\varrho}(u,t)\ du \le C(T)$$

for every $T > 0$, $\varepsilon \in (0,1]$, and $\varrho \ge 1$. As a straightforward consequence of the superlinearity (3.29) of φ and the above inequality, we realize that

$$\lim_{R\to+\infty} \sup_{t\in[0,T]} \int_R^\infty u\ f_{\varepsilon,\varrho}(u,t)\ du = 0 \tag{3.30}$$

for every $T > 0$, uniformly with respect to $\varepsilon \in (0,1]$ and $\varrho \ge 1$. One then employs (3.30) to proceed as in Step 2 and conclude that (3.27) and (3.28) hold true. We next argue as in Step 2 to show that f and f_ε satisfy (2.7) and (2.8), respectively. In addition, it readily follows from (3.13), (3.19), (3.20), and (3.30) that f_ε and f satisfy (2.10) and that the convergence (3.20) can be improved to the one claimed in (2.12). □

Proof of Proposition 2.4. We proceed along the same lines as those of Lemma 3.1. We put $P(x) = x_+$ for $x \in \mathbb{R}$ and

$$\sigma(t) = \|f_0\|_{L^\infty}\ \exp\left(\alpha\ t\ \int_0^\infty u\ f_0(u)\ du\right)$$

for $t \ge 0$. For $\varepsilon \in (0,1]$ and $\varrho \ge 1$, it follows from (1.18) and (3.10) that

$$\begin{aligned}
&\frac{d}{dt}\int_0^\infty P(f_{\varepsilon,\varrho}-\sigma)\ du \\
&\le \int_0^\infty \int_0^v \left\{ \frac{P'(f_{\varepsilon,\varrho}(v+\varepsilon w)-\sigma) - P'(f_{\varepsilon,\varrho}(v)-\sigma)}{\varepsilon} \right\} a_\varrho(v,w)\ f_{\varepsilon,\varrho}(v)\ f_{\varepsilon,\varrho}(w)\ dwdv \\
&\quad - \int_0^\infty P'(f_{\varepsilon,\varrho}-\sigma)\ \sigma'\ du\,.
\end{aligned}$$

Owing to the convexity of P, we have for $x,\ y \ge 0$,

$$\begin{aligned} x\ (P'(y-\sigma) - P'(x-\sigma)) &\le y\ P'(y-\sigma) + P(x-\sigma) - P(y-\sigma) - x\ P'(x-\sigma) \\ &\le \sigma\ (P'(y-\sigma) - P'(x-\sigma))\ . \end{aligned}$$

Consequently, since $P' \ge 0$,

$$\begin{aligned} &\int_0^\infty \int_0^v \left\{ \frac{P'(f_{\varepsilon,\varrho}(v+\varepsilon w)-\sigma) - P'(f_{\varepsilon,\varrho}(v)-\sigma)}{\varepsilon} \right\}\ a_\varrho(v,w)\ f_{\varepsilon,\varrho}(v)\ f_{\varepsilon,\varrho}(w)\ dwdv \\ &\le \sigma\ \int_0^\infty \int_0^v \left\{ \frac{P'(f_{\varepsilon,\varrho}(v+\varepsilon w)-\sigma) - P'(f_{\varepsilon,\varrho}(v)-\sigma)}{\varepsilon} \right\}\ a_\varrho(v,w)\ f_{\varepsilon,\varrho}(w)\ dwdv \\ &\le \sigma\ \int_0^\infty \int_w^\infty \left\{ \frac{a_\varrho(v-\varepsilon w,w) - a_\varrho(v,w)}{\varepsilon} \right\}\ P'(f_{\varepsilon,\varrho}(v)-\sigma)\ f_{\varepsilon,\varrho}(w)\ dvdw \\ &\le \alpha\ \sigma\ \int_0^\infty u\ f_0(u)\ du\ \int_0^\infty P'(f_{\varepsilon,\varrho}(v)-\sigma)\ dv\,, \end{aligned}$$

where we have used (2.1) and (3.13) to obtain the last inequality. Thanks to the choice of σ, combining the previous two inequalities yields that

$$\frac{d}{dt} \int_0^\infty P(f_{\varepsilon,\varrho} - \sigma)\ du \le 0\,,$$

whence

$$\|f_{\varepsilon,\varrho}(t)\|_{L^\infty} \le \sigma(t)\,, \quad t \ge 0\,.$$

Proposition 2.4 then follows from the above inequality by (3.19) and (3.20). ∎

4. Large time behavior and gelation. Throughout this section, we consider $f_0 \in X^+$, $f_0 \not\equiv 0$ and denote by f a weak solution to (1.5), (1.2) on $[0,+\infty)$. We also denote the total mass of f at time t by $M_1(t)$, that is,

$$M_1(t) = \int_0^\infty u\ f(u,t)\ du\,. \tag{4.1}$$

Since the OHS equation (1.5) accounts for only coagulation reactions, the total number of particles (which is nothing but the L^1-norm of f) is expected to decrease to zero as time increases to infinity. More precisely, we have the following result.

PROPOSITION 4.1. *For $k \in [0,1]$ and $t_2 \ge t_1 \ge 0$, we have*

$$\int_0^\infty u^k\ f(u,t_2)\ du \le \int_0^\infty u^k\ f(u,t_1)\ du\,. \tag{4.2}$$

Assume further that for each $\eta > 0$, there is $\delta_\eta > 0$ such that

$$a(u,v) \ge \delta_\eta \quad \text{for} \quad (u,v) \in (\eta,+\infty) \times (\eta,+\infty)\,. \tag{4.3}$$

Then

$$\lim_{t \to +\infty} \int_0^\infty f(u,t)\ du = 0\,.$$

Proof. Let $R \ge 1$ and take $\phi(v) = \min\{v^k, R^k\}$, $v \in \mathbb{R}_+$, in (2.7) with $t = t_1$ and $t = t_2$. Substracting the resulting identities yields

$$\int_0^\infty f(u,t_2)\ \phi(u)\ du = \int_0^\infty f(u,t_1)\ \phi(u)\ du + \int_{t_1}^{t_2}\int_0^\infty\int_0^v [w\ \phi'(v) - \phi(w)]\ a(v,w)\ f(v,s)\ f(w,s)\ dwdvds\,.$$

Since $k \in [0,1]$, we have $[w\ \phi'(v) - \phi(w)] \le 0$ for $0 \le w \le v$, from which we conclude that

$$\int_0^\infty f(u,t_2)\ \min\{u^k, R^k\}\ du \le \int_0^\infty f(u,t_1)\ \min\{u^k, R^k\}\ du$$

for every $R \ge 1$. Owing to Definition 2.1, we may let $R \to +\infty$ and deduce (4.2).

We next prove the second assertion of Proposition 4.1. For $U > 0$ and $t \ge 0$, we put

$$L(U,t) = \int_0^U f(u,t)\ du\,.$$

We fix $U > 0$ and put $\phi_\nu(v) = \min\{1, (U+\nu-v)_+/\nu\}$ for $v \in \mathbb{R}_+$ and $\nu \in (0,1)$. Then $\phi_\nu \in W^{1,\infty}(\mathbb{R}_+)$ and $\phi_\nu' \le 0$. We then infer from (2.7) with $\phi = \phi_\nu$ that for $t_2 \ge t_1 \ge 0$,

$$\begin{aligned}&\int_0^\infty (f(u,t_2) - f(u,t_1))\ \phi_\nu(u)\ du\\ &\le -\int_{t_1}^{t_2}\int_0^\infty\int_0^v \phi_\nu(w)\ a(v,w)\ f(v,s)\ f(w,s)\ dwdvds\\ &\le -\frac{1}{2}\int_{t_1}^{t_2}\int_0^U\int_0^U a(v,w)\ f(v,s)\ f(w,s)\ dwdvds\,.\end{aligned}$$

We may then let $\nu \to 0$ and deduce that

$$L(U,t_2) - L(U,t_1) \le -\frac{1}{2}\int_{t_1}^{t_2}\int_0^U\int_0^U a(v,w)\ f(v,s)\ f(w,s)\ dwdvds\,. \tag{4.4}$$

A first consequence of (4.4) is that $L(U,.)$ is a nondecreasing and nonnegative function of time, and there exists $L(U) \ge 0$ such that

$$\lim_{t\to+\infty} L(U,t) = L(U)\,. \tag{4.5}$$

As a second consequence of (4.4), we realize that (4.3) yields, for $\eta \in (0,U)$, that

$$\begin{aligned}\int_0^t \left(\int_\eta^U f(v,s)\ dv\right)^2 ds &\le \frac{1}{\delta_\eta}\int_0^t\int_0^U\int_0^U a(v,w)\ f(v,s)\ f(w,s)\ dwdvds\\ &\le \frac{2}{\delta_\eta}\ L(U,0)\,,\end{aligned}$$

and thus

$$t \mapsto L(U,t) - L(\eta,t) \in L^2(0,+\infty)\,. \tag{4.6}$$

It then readily follows from (4.5), (4.6), and the time monotonicity of $L(\eta,.)$ that

$$0 \le L(U) = L(\eta) \le L(\eta,0)\,, \quad \eta \in (0,U)\,.$$

Since $f_0 \in L^1(\mathbb{R}_+)$, we pass to the limit as $\eta \to 0$ in the above identity to conclude that $L(U) = 0$ for each $U > 0$. We next observe that (4.2) entails that

$$\|f(t)\|_{L^1} \le L(U,t) + \frac{1}{U} \int_U^\infty u\ f(u,t)\ du \le L(U,t) + \frac{1}{U} \int_0^\infty u\ f_0(u)\ du\,,$$

from which the second assertion of Proposition 4.1 readily follows by first letting $t \to +\infty$ and then $U \to +\infty$. □

We now turn to the proof of Theorem 2.5, beginning with the following lemma.

LEMMA 4.2. *For $t_2 \ge t_1 \ge 0$, we have*

$$\int_{t_1}^{t_2} \int_0^\infty \int_0^\infty a(v,w)\ f(v,s)\ f(w,s)\ dvdwds \le 2 \int_0^\infty f(v,t_1)\ dv \tag{4.7}$$

and

$$\int_{t_1}^{t_2} \int_R^\infty \int_R^\infty a(v,w)\ f(v,s)\ f(w,s)\ dvdwds \le \frac{2}{R} \int_0^\infty v\ f(v,t_1)\ dv \tag{4.8}$$

for $R > 0$.

Proof. We first take $\phi = 1$ in (2.7) to obtain (4.7). We next take $\phi(v) = \min\{v, R\}$ in (2.7) and notice that

$$\begin{aligned} &w\ \phi'(v) - \phi(w) = -R \quad \text{if} \quad v \ge w \ge R\,, \\ &w\ \phi'(v) - \phi(w) \le 0 \quad \text{if} \quad v \ge w \ge 0 \quad \text{and} \quad w \le R \end{aligned}$$

to conclude that (4.8) holds true. □

We now argue as in [9, Theorem 1.1] to complete the proof of Theorem 2.5.

Proof of Theorem 2.5. Assume first that $\lambda \in (1,2)$. We put $\zeta(v) = \left(v^{1-\lambda/2} - 1\right)_+$, $v \in \mathbb{R}_+$, and notice that

$$\mathcal{J} = \int_0^\infty \zeta'(v)\ v^{-1/2}\ dv < +\infty\,,$$

since $\lambda + 1 > 2$ and ζ vanishes in a neighborhood of $v = 0$. For $t_2 \ge t_1 \ge 0$, it follows from the Hölder inequality, (2.15), and (4.8) that

$$\begin{aligned} &\int_{t_1}^{t_2} \left(\int_0^\infty \zeta'(R) \int_R^\infty u^{\lambda/2}\ f(u,s)\ dudR \right)^2 ds \\ &\le \mathcal{J} \int_{t_1}^{t_2} \int_0^\infty \zeta'(R)\ R^{1/2} \left(\int_R^\infty u^{\lambda/2}\ f(u,s)\ du \right)^2 dRds \\ &\le \frac{\mathcal{J}}{K_0} \int_0^\infty \zeta'(R)\ R^{1/2} \int_{t_1}^{t_2} \int_R^\infty \int_R^\infty a(u,v)\ f(u,s)\ f(v,s)\ dvdudsdR \\ &\le \frac{2\,\mathcal{J}^2}{K_0}\ M_1(t_1)\,. \end{aligned}$$

However,

$$\int_0^\infty \zeta'(R) \int_R^\infty u^{\lambda/2}\ f(u,s)\ du dR = \int_0^\infty u^{\lambda/2}\ \zeta(u)\ f(u,s)\ du \\ \ge C(\lambda)\ \int_2^\infty u\ f(u,s)\ du\,.$$

Combining the previous inequalities yields

$$\int_{t_1}^{t_2} \left(\int_2^\infty u\ f(u,s)\ du \right)^2\ ds \le C\ M_1(t_1) \tag{4.9}$$

for some constant C depending only on λ and K_0. Next it follows from (2.15) and (4.7) that

$$\int_{t_1}^{t_2} \left(\int_0^2 u\ f(u,s)\ du \right)^2\ ds \le C\ \int_0^\infty\ f(v,t_1)\ dv\,. \tag{4.10}$$

We then infer from (4.2), (4.9), (4.10), and the Young inequality that

$$\int_0^\infty M_1(s)^2\ ds \le C\ \|f_0\|_X\,. \tag{4.11}$$

Recalling (4.2), we realize that the total mass M_1 is a nondecreasing and nonnegative function which also belongs to $L^2(0,+\infty)$. Therefore,

$$\lim_{t\to+\infty} M_1(t) = 0\,,$$

which readily implies that $T_{gel} < +\infty$ since $M_1(0) > 0$.

We next consider the case $\lambda = 2$. It then readily follows from (2.15) and (4.7) that (4.11) holds true, and we argue as above to complete the proof. ∎

5. Compactly supported initial data. Throughout this section, we consider $f_0 \in X^+$ and denote by f a weak solution to (1.5), (1.2) on $[0,+\infty)$. As in the previous section, we denote the total mass of f at time t by $M_1(t)$; see (4.1).

Proof of Theorem 2.6. Since

$$(u,t) \mapsto \int_0^u v\ a(u,v)\ f(v,t)\ dv$$

belongs to $\mathcal{C}([0,+\infty)\times[0,+\infty))$, the first assertion of Theorem 2.6 follows at once from the Cauchy–Péano theorem. The monotonicity of R is then a consequence of the nonnegativity of f and a. We next put

$$F(u,t) = \int_u^\infty f(v,t)\ dv\,, \quad F_0(u) = F(u,0)\,,$$
$$\lambda(u,t) = \int_0^u v\ a(u,v)\ f(v,t)\ dv\,, \quad \mu(u,t) = \int_u^\infty a(u,v)\ f(v,t)\ dv$$

for $(u,t) \in \mathbb{R}^2_+$. Then (2.2) and (4.2) ensure that $F \in L^\infty(0,+\infty;W^{1,1}(\mathbb{R}_+))$ and λ and μ belong to $L^\infty_{\rm loc}([0,+\infty)^2)$. Moreover, the boundedness of f, (2.1), and (4.2) imply that

$$\partial_u\lambda(u,t) = u\ a(u,u)\ f(u,t) + \int_0^u v\ \partial_u a(u,v)\ f(v,t)\ dv \in L^\infty_{\rm loc}([0,+\infty)^2)\,,$$

and we infer from (2.7) that F satisfies

$$\partial_t F(u,t) + \partial_u \left(\lambda(u,t)\ F(u,t)\right) = \partial_u \lambda(u,t)\ F(u,t) - \int_u^\infty \mu(v,t)\ f(v,t)\ dv$$

a.e. in $\mathbb{R}_+$ for each $t \ge 0$ with $F(0) = F_0$. The regularity of f and a then implies that $F \in W^{1,\infty}_{loc}([0,+\infty) \times \mathbb{R}_+)$. Since μ and f are nonnegative, it follows from (2.17) and the above equation that for $t_0 \in (0, T_\star)$ and $t \in [0, t_0]$, we have

$$\begin{aligned}
\frac{d}{dt} \int_{R(t)}^{R(t_0)} F(u,t)\ du &= \int_{R(t)}^{R(t_0)} \partial_t F(u,t)\ du - F(R(t),t)\ R'(t) \\
&\le -\lambda(R(t_0),t)\ F(R(t_0),t) + \lambda(R(t),t)\ F(R(t),t) \\
&\quad + \int_{R(t)}^{R(t_0)} F(u,t)\ \partial_u \lambda(u,t)\ du - F(R(t),t)\ R'(t) \\
&\le \|\partial_u \lambda\|_{L^\infty((0,R(t_0))\times(0,t_0))} \int_{R(t)}^{R(t_0)} F(u,t)\ du\,.
\end{aligned}$$

We now apply the Gronwall lemma and use (2.16) to conclude that

$$\int_{R(t)}^{R(t_0)} F(u,t)\ du \le \int_{R_0}^{R(t_0)} F_0(u)\ du\ \exp\left(t_0\ \|\partial_u \lambda\|_{L^\infty((0,R(t_0))\times(0,t_0))}\right) = 0$$

for each $t \in [0, t_0]$. Consequently,

$$F(R(t),t) = \int_{R(t)}^\infty f(u,t)\ du = 0$$

for each $t \in [0, t_0]$, whence the second assertion of Theorem 2.6 since t_0 is arbitrary in $[0, T_\star)$. ∎

Proof of Corollary 2.7. Let $t \in (0, T_\star)$. Since $f(t)$ is compactly supported, (2.19) follows at once from (2.7) with $\phi(u) = u$, $u \in \mathbb{R}_+$. ∎

We next show that $T_\star = +\infty$ for subadditive coagulation kernels.

LEMMA 5.1. *Assume that $a \in \mathcal{C}([0,+\infty)^2)$ satisfies* (2.4). *Then $T_\star = +\infty$.*

Proof. The subadditivity (2.4) of the coagulation kernel, (2.17), and (4.2) entail that for $t \in [0, T_\star)$,

$$\begin{aligned}
R'(t) &\le K_1 \int_0^{R(t)} u\ (1 + R(t) + u)\ f(u,t)\ du \\
&\le K_1\ (1 + 2\ R(t)) \int_0^\infty u\ f(u,t)\ du \\
&\le K_1\ (1 + 2\ R(t)) \int_0^\infty u\ f_0(u)\ du\,,
\end{aligned}$$

and the Gronwall lemma ensures that (2.18) cannot occur, whence $T_\star = +\infty$. ∎

We next give some bounds from above and below for R for additive and product coagulation kernels. We first consider additive coagulation kernels.

LEMMA 5.2. *Assume that*

$$a(u,v) = u^\lambda + v^\lambda\,, \quad (u,v) \in \mathbb{R}_+^2\,,$$

with $\lambda \in [0,1]$. Then $T_\star = +\infty$ and, for $t \ge 0$,

$$\left(R(0)^{1-\lambda} + (1-\lambda)\ M_1(0)\ t\right)^{1/(1-\lambda)} \le R(t) \le \left(R(0)^{1-\lambda} + 2\ (1-\lambda)\ M_1(0)\ t\right)^{1/(1-\lambda)}$$

if $\lambda \in [0,1)$ *and*

$$R(0)\ e^{M_1(0)t} \le R(t) \le R(0)\ e^{2M_1(0)t}$$

if $\lambda = 1$.

Proof. By Lemma 5.1, we have $T_\star = +\infty$, and we infer from (2.17) that

$$R'(t) = \int_0^{R(t)} u\ \left(R(t)^\lambda + u^\lambda\right)\ f(u,t)\ du\,.$$

Owing to (2.19), the above identity yields that

$$R'(t) \le 2\ R(t)^\lambda\ \int_0^\infty u\ f(u,t)\ du \le 2\ M_1(0)\ R(t)^\lambda\,,$$

whence the upper bound by the Gronwall lemma. Similarly,

$$R'(t) \ge R(t)^\lambda\ \int_0^\infty u\ f(u,t)\ du \ge M_1(0)\ R(t)^\lambda\,,$$

and the expected lower bound follows again from the Gronwall lemma. ∎

Remark 5.3. Of course, similar upper or lower bounds are also available for coagulation kernels being bounded from below or above by an additive coagulation kernel.

We next turn to product kernels, that is, $a(u,v) = (u\ v)^{\lambda/2}$, $(u,v) \in \mathbb{R}_+^2$, for some $\lambda \in [0,2]$. Observe that by Lemma 5.1, $T_\star = +\infty$ if $\lambda \in [0,1]$, while Theorem 2.5 entails that $T_\star < +\infty$ if $\lambda \in (1,2]$. The estimates we obtain for product kernels will therefore be of a different nature according to whether $\lambda \in [0,1]$ or $\lambda \in (1,2]$.

LEMMA 5.4. *Assume that*

$$a(u,v) = (u\ v)^{\lambda/2}\,, \quad (u,v) \in \mathbb{R}_+^2\,,$$

with $\lambda \in [0,2]$.

1. *If* $\lambda \in [0,1]$, *then* $T_\star = +\infty$ *and*

$$R(t) \le \left(R(0)^{1-\lambda} + (1-\lambda)\ M_1(0)\ t\right)^{1/(1-\lambda)}$$

if $\lambda \in [0,1)$ *and*

$$R(t) \le R(0)\ e^{M_1(0)t}$$

if $\lambda = 1$.

2. *If* $\lambda \in (1,2]$, *then* $T_\star < +\infty$ *and*

$$T_\star \ge \frac{1}{(\lambda-1)\ M_1(0)\ R(0)^{\lambda-1}}\,.$$

Proof. By (2.17) and (4.2), we have

$$R'(t) = R(t)^{\lambda/2}\ \int_0^{R(t)} u^{1+\lambda/2}\ f(u,t)\ du \le M_1(0)\ R(t)^\lambda$$

for $t \in [0,T_\star)$. Lemma 5.4 then readily follows by direct integration of the previous inequality. ∎

An interesting question is whether it is possible to obtain similar estimates from below for R when $\lambda \in [0,1]$ and from above for $T_\star$ when $\lambda \in (1,2]$, but this seems to be less obvious.

Acknowledgments. Part of this work was performed while the second author enjoyed the hospitality and support of the Institute of Applied Mathematics and Mechanics, Warsaw University.

REFERENCES

[1] N. Bellomo and M. Lo Schiavo, *Lecture Notes on the Mathematical Theory of Generalized Boltzmann Models*, World Scientific, River Edge, NJ, 2000.

[2] F. P. da Costa, *On the positivity of solutions to the Smoluchowski equations*, Mathematika, 42 (1995), pp. 406–412.

[3] S. Chandrasekhar, *Stochastic problems in physics and astronomy*, Rev. Modern Phys., 15 (1943), pp. 1–91.

[4] C. Dellacherie and P. A. Meyer, *Probabilités et Potentiel, Chapitres* I *à* IV, Hermann, Paris, 1975.

[5] P. G. J. van Dongen and M. H. Ernst, *Scaling solutions of Smoluchowski's coagulation equation*, J. Statist. Phys., 50 (1988), pp. 295–329.

[6] R. L. Drake, *A general mathematical survey of the coagulation equation*, in Topics in Current Aerosol Research, Part 2, International Reviews in Aerosol Physics and Chemistry, Pergamon Press, Oxford, UK, 1972, pp. 203–376.

[7] P. B. Dubovski, *A "triangle" of interconnected coagulation models*, J. Phys. A, 32 (1999), pp. 781–793.

[8] P. B. Dubovski, *A new discrete model of coagulation kinetics and the properties of its continuous analogue*, Mat. Model., 12 (2000), pp. 4–15 (in Russian).

[9] M. Escobedo, S. Mischler, and B. Perthame, *Gelation in coagulation and fragmentation models*, Comm. Math. Phys., 231 (2002), pp. 157–188.

[10] I. Jeon, *Existence of gelling solutions for coagulation-fragmentation equations*, Comm. Math. Phys., 194 (1998), pp. 541–567.

[11] M. Lachowicz and D. Wrzosek, *Nonlocal bilinear equations. Equilibrium solutions and diffusive limit*, Math. Models Methods Appl. Sci., 11 (2001), pp. 1393–1409.

[12] Ph. Laurençot, *The Lifshitz-Slyozov equation with encounters*, Math. Models Methods Appl. Sci., 11 (2001), pp. 731–748.

[13] Ph. Laurençot and S. Mischler, *From the discrete to the continuous coagulation-fragmentation equations*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 1219–1248.

[14] Lê Châu-Hoàn, *Etude de la classe des opérateurs m-accrétifs de* $L^1(\Omega)$ *et accrétifs dans* $L^\infty(\Omega)$, Thèse de $3^{\text{ème}}$ cycle, Université de Paris VI, 1977.

[15] J. H. Oort and H. C. van de Hulst, *Gas and smoke in interstellar space*, Bull. Astron. Inst. Netherland, 10 (1946), pp. 187–210.

[16] A. S. Perelson and R. W. Samsel, *Kinetics of red blood cell aggregation: An example of geometric polymerization*, in Kinetics of Aggregation and Gelation, F. Family and D. P. Landau, eds., Elsevier, Amsterdam, 1984, pp. 137–144.

[17] V. S. Safronov, *Evolution of the Protoplanetary Cloud and Formation of the Earth and the Planets*, Israel Program for Scientific Translations, Jerusalem, 1972.

[18] M. Smoluchowski, *Drei Vorträge über Diffusion, Brownsche Molekularbewegung und Koagulation von Kolloidteilchen*, Physik. Zeitschr., 17 (1916), pp. 557–599.

[19] M. Smoluchowski, *Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Lösungen*, Zeitschrift f. physik. Chemie, 92 (1917), pp. 129–168.

[20] I. W. Stewart, *A global existence theorem for the general coagulation-fragmentation equation with unbounded kernels*, Math. Methods Appl. Sci., 11 (1989), pp. 627–648.

[21] I. I. Vrabie, *Compactness Methods for Nonlinear Evolutions*, 2nd ed., Pitman Monogr. Surveys Pure Appl. Math. 75, Longman Scientific and Technical, Harlow, UK, 1995.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1422–1434

FIRST BOUNDARY VALUE PROBLEM FOR THE DIFFUSION EQUATION I. ITERATED LOGARITHM TEST FOR THE BOUNDARY REGULARITY AND SOLVABILITY*

UGUR G. ABDULLA†

Abstract. This paper establishes a precise sufficient condition for the regularity of a boundary point of an arbitrary open subset of $\mathbb{R}^{N+1}$ $(N \geq 2)$ and for the solvability of the first boundary value problem for the diffusion (or heat) equation in general domains.

Key words. first boundary value problem, diffusion (or heat) equation, boundary regularity, iterated logarithm test

AMS subject classifications. 35K05, 60J65

PII. S0036141002415049

1. Introduction. Let $\Omega \subset \mathbb{R}^{N+1}$ $(N \geq 2)$ denote any bounded open subset and $\partial\Omega$ its topological boundary. We write a typical point as $z = (x,t) = (x_1, \overline{x}, t), x = (x_1, \overline{x}) \in \mathbb{R}^N, \overline{x} = (x_2, \dots, x_N) \in \mathbb{R}^{N-1}, t \in \mathbb{R}$. For a given point $z_0 = (x^0, t_0)$ and a positive number ϵ define the cylinder

$$Q(z_0, \epsilon) = \{z : |x - x_0| < \epsilon, t_0 - \epsilon < t < t_0\}.$$

We split $\partial\Omega$ as $\partial\Omega = \mathcal{P}\Omega \bigcup \mathcal{D}\Omega$, where $\mathcal{P}\Omega$ is the set of all points $z_0 \in \partial\Omega$ such that for any $\epsilon > 0$, the cylinder $Q(z_0, \epsilon)$ contains points not in Ω. The set $\mathcal{P}\Omega$ is called the parabolic boundary of Ω. The set $\mathcal{D}\Omega$ is naturally called the top boundary of Ω. We split also $\mathcal{P}\Omega$ as $\mathcal{P}\Omega = \mathcal{S}\Omega \bigcup \mathcal{B}\Omega$, where $\mathcal{B}\Omega$ is the set of all points $z_0 \in \mathcal{P}\Omega$ such that for some $\epsilon > 0$, the cylinder $Q(z_0, \epsilon)$ lies outside Ω. The set $\mathcal{B}\Omega$ is naturally called the bottom boundary of Ω, while $\mathcal{S}\Omega$ will be called the lateral boundary of Ω.

For $u \in C^{2,1}_{x,t}(\Omega)$, we define the diffusion (or heat) operator

$$\mathbf{D}u = u_t - \Delta u = u_t - \sum_{i=1}^{N} u_{x_i x_i}, \qquad z \in \Omega.$$

A function $u \in C^{2,1}_{x,t}(\Omega)$ is called parabolic in Ω if $\mathbf{D}u = 0$ for $z \in \Omega$. Let $f : \mathcal{P}\Omega \to \mathbb{R}$ be a bounded function. The first boundary value problem (FBVP) may be formulated as follows:

Find a function u which is parabolic in Ω and satisfies the conditions

$$f_* \leq u_* \leq u^* \leq f^* \qquad \text{for } z \in \mathcal{P}\Omega, \tag{1.1}$$

where f_*, u_* (or f^*, u^*) are lower (or upper) limit functions of f and u, respectively.

In particular, if $f \in C(\mathcal{P}\Omega; \mathbb{R})$, from (1.1) it follows that u takes continuously the given values of f on $\mathcal{P}\Omega$. The strategy for solving the FBVP may be well expressed by the citation from the classical paper [W] on the Dirichlet problem (DP) for the

*Received by the editors September 23, 2002; accepted for publication (in revised form) December 9, 2002; published electronically May 15, 2003.
http://www.siam.org/journals/sima/34-6/41504.html
†Max-Planck Institute for Mathematics in the Sciences, Leipzig 04103, Germany (abdulla@math.uni-leipzig.de). Current address: School of Mathematics and Computer Science, University of Leipzig, Augustusplatz 10/11, Leipzig 04109, Germany.

Laplace equation. As pointed out by Lebesgue and independently by Wiener [W], "the DP divides itself into two parts, the first of which is the determination of a harmonic function corresponding to certain boundary conditions, while the second is the investigation of the behaviour of this function in the neighbourhood of the boundary." The same strategy, obviously replacing harmonic function with parabolic function and boundary with the parabolic boundary, is applicable to the FBVP for the diffusion equation. As in the case of the Laplace equation, a generalized solution to the FBVP for the diffusion equation may be constructed by Perron's super- or subsolutions method (see section 2). However, in general the generalized solution doesn't satisfy (1.1).

We say that a point $z_0 \in \mathcal{P}\Omega$ is regular if for any bounded function $f : \mathcal{P}\Omega \to \mathbb{R}$, the generalized solution of the FBVP constructed by Perron's method satisfies (1.1) at the point z_0. It is well known that the boundary points $z_0 \in \mathcal{B}\Omega$ are always regular.

The principal result of this paper is the characterization of the regularity of the boundary points $z_0 \in \mathcal{S}\Omega$ via local geometry of the lateral boundary near this point.

Consider the following domains:

$$\mathcal{G}_\rho^1 = \{z : x_1^2 < 4\xi \log \rho(\xi), (\overline{x}, t) \in P(\delta)\},$$

$$\mathcal{G}_\rho^2 = \{z : -2(\xi \log \rho(\xi))^{\frac{1}{2}} < x_1 < 2(-\delta \log \rho(-\delta))^{\frac{1}{2}}, (\overline{x}, t) \in P(\delta)\},$$

$$P(\delta) = \{(\overline{x}, t) : -\delta < \xi < 0, -\delta < \alpha t < 0\},$$

where $\delta > 0$ is a sufficiently small positive number, $\xi = \alpha t - \beta|\overline{x}|^2$, and α and β are given positive numbers. Throughout this paper we assume that $\rho = \rho(\xi), -\delta \le \xi < 0$ is a positive and continuously differentiable function satisfying the following condition:

$$\rho(\xi) \downarrow 0, \quad \xi\rho^{-1}(\xi)\rho'(\xi) \to 0 \qquad \text{as } \xi \uparrow 0. \tag{1.2}$$

Applying l'Hôpital's rule to (1.2), it follows that $\xi \log \rho(\xi) \to 0$ as $\xi \uparrow 0$. In Figures 1 and 2 the domains $\mathcal{G}_\rho^1$ and $\mathcal{G}_\rho^2$ are described when $N = 2$. The parabolic boundary $\mathcal{P}\mathcal{G}_\rho^1$ consists of two manifolds

$$\mathcal{L}_\pm = \{z \in \overline{\mathcal{G}_\rho^1} : x_1 = \pm 2(\xi \log \rho(\xi))^{\frac{1}{2}}, t > -\delta\alpha^{-1}\}$$

and the cylindrical hypersurface

$$\{z \in \overline{\mathcal{G}_\rho^1} : \xi = -\delta, t > -\delta\alpha^{-1}\}.$$

The bottom boundary $\mathcal{B}\mathcal{G}_\rho^1$ consists of a line segment $\{z : x_1^2 < -4\delta \log \rho(-\delta), |\overline{x}| = 0, t = -\delta\alpha^{-1}\}$. The parabolic boundary $\mathcal{P}\mathcal{G}_\rho^2$ differs from $\mathcal{P}\mathcal{G}_\rho^1$ by replacing the manifold $\mathcal{L}_+$ with

$$\{z \in \overline{\mathcal{G}_\rho^2} : x_1 = 2(-\delta \log \rho(-\delta))^{\frac{1}{2}}, (\overline{x}, t) \in \overline{P(\delta)}\}.$$

Our main theorem reads as follows.

THEOREM 1.1. *Let* $\alpha + 2(N-1)\beta \le 1$ *and*

$$\lim_{\epsilon \uparrow 0} \int_{-\delta}^{\epsilon} \frac{\rho(\eta)}{\eta} d\eta = -\infty. \tag{1.3}$$

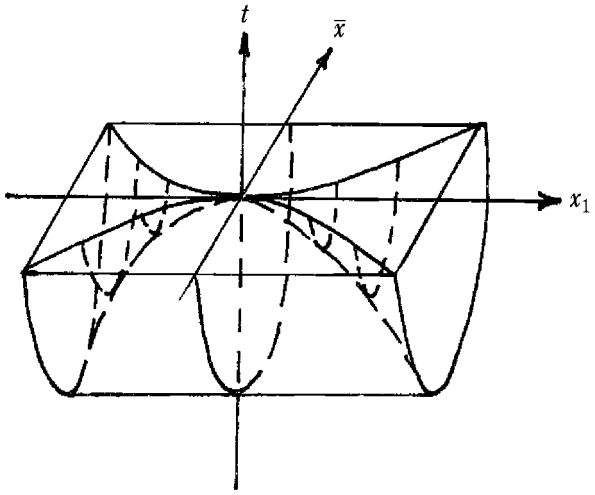


FIG. 1. *The domain* $\mathcal{G}_\rho^1$ *when* $N = 2$.

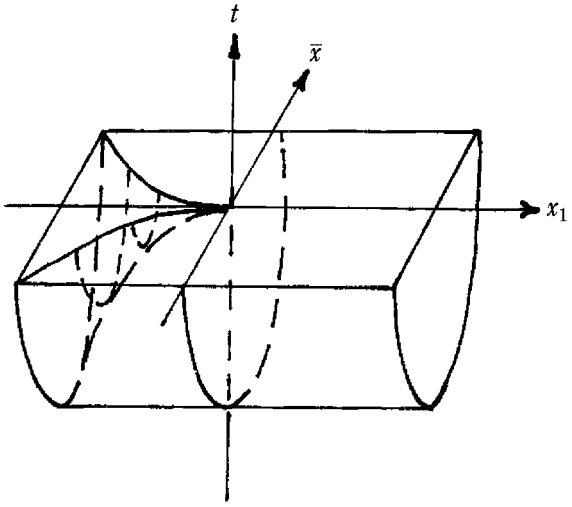


FIG. 2. *The domain* $\mathcal{G}_\rho^2$ *when* $N = 2$.

Then the origin ($\mathcal{O}$) *is a regular point for* $\mathcal{G}_\rho^1$ *(or* $\mathcal{G}_\rho^2$*) and the FBVP is solvable in* $\mathcal{G}_\rho^1$ *(or* $\mathcal{G}_\rho^2$*).*

Some examples of functions ρ that satisfy (1.2), (1.3) are

$$\rho(\xi) = |\log|\xi||^{-1}, \quad \rho(\xi) = \left\{ |\log|\xi|| \prod_{k=2}^{n} \log_k|\xi| \right\}^{-1}, \qquad n = 2, 3, \ldots, \tag{1.4}$$

where we use the following notation:

$$\log_2|\xi| = \log|\log|\xi||, \quad \log_n|\xi| = \log\log_{n-1}|\xi|, \qquad n \geq 3.$$

Theorem 1.1 provides a general sufficient condition for the regularity of the boundary points $z_0 \in \mathcal{S}\Omega$ and for the solvability of the FBVP in Ω.

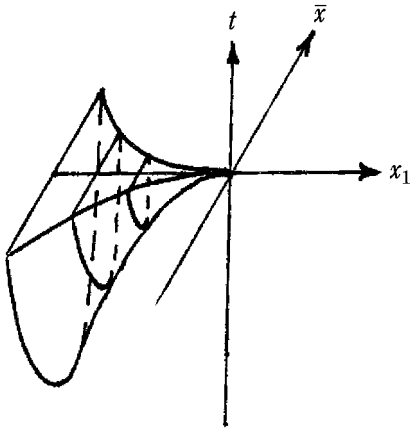


FIG. 3. *The domain $\mathcal{A}_\rho$ when $N = 2$.*

Let $\mathcal{A}_\rho = \mathcal{G}\backslash\overline{\mathcal{G}_\rho^2}$, where

$$\mathcal{G} = \{z : x_1^2 < -4\delta \log \rho(-\delta), (\overline{x}, t) \in P(\delta)\}.$$

In Figure 3 the domain $\mathcal{A}_\rho$ is described when $N = 2$. We call the origin the vertex of $\mathcal{A}_\rho$. Consider the rigid body displacements of $\mathcal{A}_\rho$ composed of translations and (or) rotations in x-space and shift along the t-axis.

DEFINITION 1.2. *We shall say that Ω satisfies the exterior $\mathcal{A}_\rho$-condition at the point $z_0 \in \mathcal{S}\Omega$ if after the above-mentioned displacement the vertex of $\mathcal{A}_\rho$ coincides with z_0 and for all sufficiently small δ, $\mathcal{A}_\rho$ lies in the exterior of Ω.*

Theorem 1.1 implies the following more general result.

THEOREM 1.3. *The boundary point $z_0 \in \mathcal{S}\Omega$ is regular if Ω satisfies the exterior $\mathcal{A}_\rho$-condition at this point. The FBVP is solvable in a region Ω which satisfies the exterior $\mathcal{A}_\rho$-condition at every point $z_0 \in \mathcal{S}\Omega$.*

In the case when the lateral boundary is locally a continuous graph, the exterior $\mathcal{A}_\rho$-condition may be expressed in terms of modulus of lower semicontinuity of the lateral boundary manifold. To make this precise, assume that for $z_0 = (x^0, t_0) \in \mathcal{S}\Omega$ there exists $\epsilon > 0$ and a continuous function ϕ such that, after a suitable rotation of x-axes, we have

$$\overline{\mathcal{S}\Omega} \cap Q(z_0, \epsilon) = \{z \in Q(z_0, \epsilon) : x_1 = \phi(\overline{x}, t)\}, \tag{1.5}$$

$$\text{sign } (x_1 - \phi(\overline{x}, t)) = 1 \qquad \text{for } z \in Q(z_0, \epsilon) \cap \Omega. \tag{1.6}$$

The exterior $\mathcal{A}_\rho$-condition is equivalent to the following one-side inequality for the function ϕ:

$$\phi(\overline{x}^0, t_0) - \phi(\overline{x}, t) \le 2(\xi' \log \rho(\xi'))^{\frac{1}{2}} \qquad \text{for } (\overline{x}, t) \in \overline{P'(\delta)}, \tag{1.7}$$

where $\delta > 0$ is a sufficiently small number, $\xi' = \alpha(t - t_0) - \beta|\overline{x} - \overline{x}^0|^2$; the domain $P'(\delta)$ coincides with $P(\delta)$ by replacing ξ with ξ' and t with $t - t_0$; the numbers $\alpha > 0, \beta > 0$ and the function ρ satisfy the conditions of Theorem 1.1. The equivalence easily follows from the fact that after the displacement according to the exterior $\mathcal{A}_\rho$-condition,

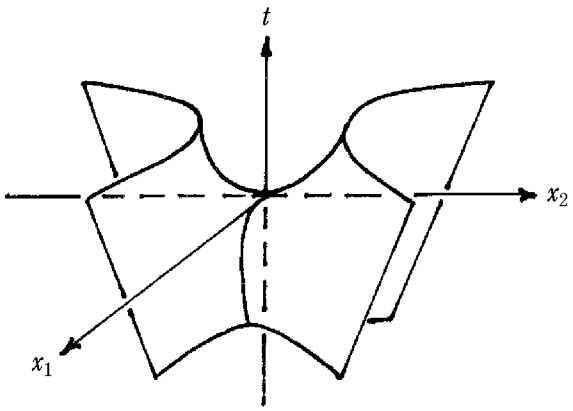


FIG. 4. *Hyperbolic paraboloid* $x_1^2 = M(-t + x_2^2)$.

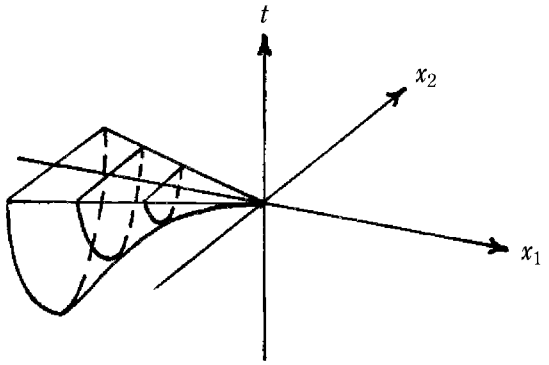


FIG. 5. *The domain* $\mathcal{M}_\delta$.

the boundary manifold which is a common boundary of the translated domains $\mathcal{A}_\rho$ and $\mathcal{G}_\rho^1$ has a representation $x_1 = \phi(\overline{x}, t), (\overline{x}, t) \in P'(\delta)$, where ϕ satisfies (1.7) with "=" instead of "$\leq$". Inequality (1.7) means that at the point $z_0 = (x^0, t_0) \in \mathcal{S}\Omega$, the lateral boundary manifold is allowed to be "slightly worse" than lower Lipschitz in the x-direction and "slightly worse" than lower $\frac{1}{2}$-Hölder in $-t$-direction. "Slightly worse" means that the related Lipschitz (or, respectively, Hölder) coefficient may converge to infinity as $(\overline{x}, t) \to (\overline{x}^0, t_0)$ but not faster than $2(-\beta \log \rho(-\beta|\overline{x} - \overline{x}^0|^2))^{\frac{1}{2}}$ (or, respectively, $2(-\alpha \log \rho(\alpha(t - t_0)))^{\frac{1}{2}}$). In the particular case when both coefficients are constant, we get the parabolic analogue of the well-known exterior cone condition for the Laplace equation. Let us formulate this condition geometrically in the spirit of our exterior $\mathcal{A}_\rho$-condition. For simplicity take $N = 2$ and consider the hyperbolic paraboloid (Figure 4)

$$x_1^2 = M(-t + x_2^2), \qquad M > 0.$$

Let $\delta > 0$ be given and consider the subsurface of the hyperbolic paraboloid which is situated in the half space $\{t \leq 0\}$ between two planes $\{x_1 = 0\}$ and $\{x_1 = -\delta^{\frac{1}{2}}\}$ (Figure 5). Consider the open domain $\mathcal{M}_\delta$ which is bounded by this subsurface and by the planes $\{t = 0\}$ and $\{x_1 = -\delta^{\frac{1}{2}}\}$ (Figure 5). We call the origin the vertex

of $\mathcal{M}_\delta$. Consider the rigid body displacements of $\mathcal{M}_\delta$ composed of translations and (or) rotations in x-space and shift along the t-axis. If after such a displacement the vertex of $\mathcal{M}_\delta$ coincides with the point $z_0 \in \mathcal{S}\Omega$, and for all sufficiently small δ, $\mathcal{M}_\delta$ lies in the exterior of Ω, then z_0 is a regular point. This fact is an easy consequence of Theorem 1.3. A similar condition is obviously true when $N > 2$ just by replacing x_2^2 with $|\overline{x}|^2$. This is exactly the parabolic analogue of the exterior cone condition for the Laplace equation, and it is natural to call it an "exterior hyperbolic paraboloid condition."

It should be mentioned that the boundary regularity result of Theorem 1.1 has a probabilistic meaning in the context of the short-time behavior of the Brownian motion trajectories for the high-dimensional diffusion processes. Without going into details, let us just formulate the probabilistic analogue of this result taking the simplest example $\rho(\xi) = |\log|\xi||^{-1}$. Consider the standard N-dimensional Brownian motion in which the coordinates of the sample path are standard one-dimensional Brownian motions. The intuitive meaning of Theorem 1.1 is that the Brownian path that starts at the origin (assuming that the process goes in the $-t$-direction) with probability 1 will reach the exterior of $\mathcal{G}_\rho^1$ within arbitrarily short time. From the classical iterated logarithm law it easily follows that with probability 1 the same trajectory will remain in the domain $\mathcal{G}_\rho^1$ within some positive time if $\alpha > 1, \beta > 0$. We explain this fact below in section 4 (see five lines after (4.1)). An important open problem is whether the same is true if $0 < \alpha \leq 1$ but $\alpha + 2(N-1)\beta > 1$. The related open problem in the context of the FBVP consists in the derivation of the precise sufficient condition for the irregularity of the boundary points. We address this issue in the next paper.

We prove the main theorems in section 3, after some preliminaries in section 2. Section 4 contains some final remarks.

Historical comments. In 1935, Petrovsky [P] presented complete results on the FBVP for the one-dimensional diffusion equation $u_t = u_{xx}$ in a plane domain whose lateral boundary is given by two continuous curves $x = \phi_1(t)$ and $x = \phi_2(t)$. Petrovsky's paper was motivated by the proof of the so-called Kolmogorov test for the distinction between the upper and the lower functions of the one-dimensional space-time Brownian motion trajectories (see [IM]). If we take $N = 1$, then our Theorem 1.1 coincides with the regularity result from [P, section 2]. Moreover, the analogue of our domain $\mathcal{G}_\rho^1$ is a plane domain between the curve $x_1^2 = 4t \log \rho(t)$ and the line $t = -\delta < 0$. We get a similar domain intersecting $\mathcal{G}_\rho^1$ with the hyperplane $\{\overline{x} = 0\}$. As was proved in [P, section 3], even for the particular example $\rho(t) = |\log|t||^{-1}$ the result is close to being an optimal in the sense that the origin is an irregular point if we replace the boundary curve with $x_1^2 = -4\alpha t \log_2|t|, \alpha > 1$. In the context of the one-dimensional Brownian motion this result repeats Khinchin's iterated logarithm law. As a direct implication of the one-dimensional results, in [P, section 4] the case $N = 2$ was also briefly considered. It was shown that the origin is an irregular point for the bounded domain lying beneath the plane $\{t = 0\}$ and bounded on its sides by the surface of revolution

$$x^2 + y^2 = -4(1+\epsilon)t \log_2|t|,$$

where $\epsilon > 0$ is an arbitrary small number. From another side, from the regularity condition formulated in [P, section 4], it follows that the origin is a regular point for the same domain if we replace the surface of revolution with the following one:

$$x^2 + y^2 = 4t \log \rho(t),$$

where $\rho(t)$ satisfies the conditions of [P, section 4] (or our Theorem 1.1). Both results are true for the case $N \geq 3$ as well. The meaning of these results is that the conditions for the regularity or irregularity of the boundary point on the top of the radially symmetric domain formed with the rotation of the plane curve $x_1^2 = 4t \log \rho(t)$ around the t-axis are the same as in the one-dimensional case. Probabilistic interpretation of this fact may be given in the context of the multidimensional Brownian motion.

Starting from 1954–1955, Wiener's ideas [W] concerning the regularity of the boundary points for the Laplace equation were adapted to the case of the diffusion equation. In [L], Wiener-type necessary and sufficient conditions for the regularity of the boundary points in the FBVP for the diffusion equation were published. The analogue of Wiener's condition, namely, a necessary and sufficient condition which is a quasi-geometric characterization for a boundary point of an arbitrary bounded open subset of $\mathbb{R}^{N+1}$ to be regular for the diffusion equation, was established in [EG]. However, it should be mentioned that Wiener's criterion does not resolve the natural geometric and analytic question which we impose in this paper. Despite its generality, it seems impossible to derive even Petrovsky's one-dimensional results from Wiener's condition.

Another sufficient condition for the regularity of the boundary points in the FBVP for the diffusion equation, the so-called exterior tusk condition which is an analogue of the exterior cone condition for the Laplace equation, was established in [EK]. It should be pointed out that the origin does not satisfy the exterior tusk condition from [EK] as a boundary point of $\mathcal{G}_\rho^1$ (or $\mathcal{G}_\rho^2$). The exterior tusk condition is satisfied for the singularities in the $-t$-direction which are not "more flat" than a branch of parabola near its vertex. Otherwise speaking, the lower Hölder condition with Hölder exponent $\frac{1}{2}$ should be satisfied, provided that (1.5), (1.6) are valid. Similarly, under the assumptions (1.5), (1.6), the exterior tusk condition may be satisfied for the singularities in the x-direction which satisfy the exterior cone condition (or lower Lipschitz condition). For example, the exterior tusk condition is not satisfied for the origin as a boundary point of the cylindrical domain whose projection to the hyperplane $t = 0$ coincides with $\overline{\mathcal{G}_\rho^2} \cap \{t = 0\}$. Hence, the exterior tusk condition is similar to the exterior hyperbolic paraboloid condition.

2. Preliminary results. In this section we present some facts about Perron's solution of the FBVP. Lemma 2.1 is standard and demonstrates the role of barriers for the regularity of the origin for $\mathcal{G}_\rho^1$ or $\mathcal{G}_\rho^2$. Lemma 2.2 proves the equivalence of the regularity (or irregularity) of the origin for $\mathcal{G}_\rho^1$ and $\mathcal{G}_\rho^2$, which allows us to prove Theorem 1.1 only for $\mathcal{G}_\rho^1$.

It should be mentioned that the results of this section are general, and we do not need to assume that the conditions of Theorem 1.1 and the second condition from (1.2) are satisfied. However, we need to assume that $\xi \log \rho(\xi) \to 0$ as $\xi \uparrow 0$.

A bounded open subset $U \in \mathbb{R}^{N+1}$ is called regular if for each continuous function $\phi \in C(\partial U; \mathbb{R})$ there exists one (and only one) function H_ϕ^U, which is parabolic in U, and

$$\lim_{z \to z_0, z \in U} H_\phi^U = \phi(z_0) \qquad \text{for all } z_0 \in \partial U.$$

A function $u \in C(\Omega)$ is called superparabolic in Ω if the following conditions are satisfied:

(a) u is lower semicontinuous; $-\infty < u \leq +\infty$, $u < +\infty$ on a dense subset of Ω;
(b) if $U \subset \overline{U} \subset \Omega$ is a regular open set, $\phi \in C(\partial U; \mathbb{R})$, and $\phi \leq u$ on ∂U, then $H_\phi^U \leq u$ in U.

A function v is called a subparabolic if $-v$ is a superparabolic. For example, any function $u \in C^{2,1}_{x,t}(\Omega)$ satisfying $\mathbf{D}u \geq 0$ (or $\mathbf{D}u \leq 0$) for $z \in \Omega$ is superparabolic (or subparabolic). The classical theory defines Perron's solution of the FBVP to be, for each $z \in \Omega$,

$$H_f^\Omega = \inf\{u(z)\},$$

where the infimum is taken over all superparabolic functions u in Ω such that

$$u_*(z_0) \geq f^*(z_0) \qquad \text{for all } z_0 \in \mathcal{P}\Omega.$$

According to the classical theory (see, for example, [D, B] for the most general framework) H_f^Ω is parabolic in Ω. However, in general it does not satisfy (1.1). A boundary point $z_0 \in \mathcal{P}\Omega$ is called regular if for arbitrary bounded boundary function f, H_f^Ω satisfies (1.1) at this point. It is well known that bottom boundary points $z_0 \in \mathcal{B}\Omega$ are always regular (see, for example, [P]). It is a standard fact in the classical theory that the boundary point $z_0 \in \mathcal{S}\Omega$ is regular if there exists a so-called regularity barrier $\overline{u}$ with the following properties:

(a) $\overline{u}$ is superparabolic in $U = Q(z_0, \epsilon) \cap \Omega$ for some $\epsilon > 0$;

(b) $\overline{u}$ is continuous and nonnegative in $\overline{U}$, vanishing only at z_0.

In particular, concerning the regularity of the origin for $\mathcal{G}_\rho^1$ or $\mathcal{G}_\rho^2$ we have the following.

LEMMA 2.1. *The origin ($\mathcal{O}$) is regular for $\mathcal{G}_\rho^1$ (or $\mathcal{G}_\rho^2$) if and only if there exists a regularity barrier $\overline{u}$ for $\mathcal{O}$ regarded as a boundary point of $\mathcal{G}_\rho^1$ (or $\mathcal{G}_\rho^2$) for sufficiently small δ.*

Proof. The proof of the "if" part is standard. To prove the "only if" part, take $f = -t + |x|^2$ and let $\overline{u} = H_f^{\mathcal{G}_\rho^1}$ be Perron's solution. Since $\rho(\xi)$ is C^1 for $\xi < 0$, from the classical theory it follows that all the boundary points $z_0 \in \mathcal{P}\mathcal{G}_\rho^1, z_0 \neq \mathcal{O}$, are regular points. But $\mathcal{O}$ is regular by our assumption. Therefore, $\overline{u} \in C(\overline{\mathcal{G}_\rho^1})$. From the strong maximum principle it follows that $\overline{u}$ is nonnegative in $\overline{\mathcal{G}_\rho^1}$ and vanishes only at $\mathcal{O}$. Thus $\overline{u}$ is a regularity barrier for $\mathcal{O}$. The proof for the domain $\mathcal{G}_\rho^2$ is similar. The lemma is proved.

The next lemma is the high-dimensional analogue of Theorem III from [P, p. 389].

LEMMA 2.2. *The origin is simultaneously regular or irregular for $\mathcal{G}_\rho^1$ and $\mathcal{G}_\rho^2$.*

Proof. Assume that $\mathcal{O}$ is regular for $\mathcal{G}_\rho^2$. Then by Lemma 2.1 there exists a regularity barrier for $\mathcal{O}$ regarded as a boundary point of $\mathcal{G}_\rho^2$. Obviously, it will also be a regularity barrier for $\mathcal{O}$ regarded as a boundary point of $\mathcal{G}_\rho^1$. From Lemma 2.1 it follows that $\mathcal{O}$ is regular for $\mathcal{G}_\rho^1$.

Conversely, assume now that $\mathcal{O}$ is regular for $\mathcal{G}_\rho^1$. Let $u = H_f^{\mathcal{G}_\rho^2}$, where $f = -t + |x|^2$. Since $\rho(\xi)$ is C^1 for $\xi < 0$, all the boundary points $z_0 \in \mathcal{P}\mathcal{G}_\rho^2, z_0 \neq \mathcal{O}$, are regular points. Accordingly, $u \in C(\overline{\mathcal{G}_\rho^2} \backslash \{\mathcal{O}\})$. Denote

$$L = \limsup_{z \to \mathcal{O}, x_1 = \phi(\overline{x}, t)} u,$$

where $\phi(\overline{x}, t) = 4\xi \log \rho(\xi)$. Obviously, we have $0 \leq L < +\infty$. Let f_1 be an arbitrary function which is defined and continuous in $\mathcal{P}\mathcal{G}_\rho^1 \backslash \{\mathcal{O}\}$, satisfying

$$f_1(x_1, \overline{x}, t) = -f_1(-x_1, \overline{x}, t) \tag{2.1}$$

and

$$\lim_{z \to \mathcal{O}, x_1 = \phi(\overline{x}, t)} f_1 = \frac{L}{2}, \qquad \lim_{z \to \mathcal{O}, x_1 = -\phi(\overline{x}, t)} f_1 = -\frac{L}{2}.$$

Choose a function f_2 in such a way that

$$f_1 + f_2 = u \qquad \text{for } z \in \mathcal{P}\mathcal{G}_\rho^1 \cap \mathcal{G}_\rho^2,$$

$$f_1 + f_2 = f \qquad \text{on the rest of } \mathcal{P}\mathcal{G}_\rho^1.$$

Since all the points $z_0 \in \mathcal{P}\mathcal{G}_\rho^1$ are regular points, Perron's solutions $u_1 = H_{f_1}^{\mathcal{G}_\rho^1}$, $u_2 = H_{f_2}^{\mathcal{G}_\rho^1}$, and $H_{f_1+f_2}^{\mathcal{G}_\rho^1}$ are continuous functions in $\overline{\mathcal{G}_\rho^1}\backslash\{\mathcal{O}\}$. Applying the maximum principle in $\mathcal{G}_\rho^1 \cap \{z : t \le -\epsilon\}$ for arbitrary sufficiently small $\epsilon > 0$ and passing to the limit as $\epsilon \downarrow 0$, we easily derive that

$$u = H_{f_1+f_2}^{\mathcal{G}_\rho^1} = H_{f_1}^{\mathcal{G}_\rho^1} + H_{f_2}^{\mathcal{G}_\rho^1} = u_1 + u_2. \tag{2.2}$$

Applying the same arguments, from (2.1) we deduce that

$$u_1(x_1, \overline{x}, t) = -u_1(-x_1, \overline{x}, t),$$

and hence

$$u_1(0, \overline{x}, t) = 0. \tag{2.3}$$

We have

$$\limsup_{z \to \mathcal{O}, x_1 = \pm\phi(\overline{x}, t)} f_2 \le \frac{L}{2}.$$

Since $\mathcal{O}$ is a regular point for $\mathcal{G}_\rho^1$, it follows that

$$\limsup_{\xi \to 0} u_2(0, \overline{x}, t) \le \frac{L}{2}. \tag{2.4}$$

From (2.2)–(2.4) we have

$$\limsup_{\xi \to 0} u(0, \overline{x}, t) \le \frac{L}{2}.$$

Since $\mathcal{O}$ is a regular point regarded as a boundary point of $\mathcal{G}_3 = \{z \in \mathcal{G}_\rho^2 : x_1 > 0\}$, we have $L \le L/2$, which implies that $L = 0$. Thus u is continuous in $\overline{\mathcal{G}_\rho^2}$ and by the strong maximum principle vanishes only at $\mathcal{O}$. Hence, u is a regularity barrier for $\mathcal{O}$ regarded as a boundary point of $\mathcal{G}_\rho^2$. From Lemma 2.1 it follows that $\mathcal{O}$ is a regular point for $\mathcal{G}_\rho^2$ as well. The lemma is proved.

COROLLARY 2.3. *Let $\mathcal{G}_0$ be a given open set in $\mathbb{R}^{N+1}$ and $\mathcal{O} \in \partial\mathcal{G}_0$, $\mathcal{G}_0^- \ne \emptyset$, where $\mathcal{G}_0^- = \{z \in \mathcal{G}_0 : t < 0\}$. If $\mathcal{G}_0^- \subset \mathcal{G}_\rho^2$, then from the regularity of $\mathcal{O}$ for $\mathcal{G}_\rho^1$ or $\mathcal{G}_\rho^2$ it follows that $\mathcal{O}$ is regular for $\mathcal{G}_0$. Otherwise speaking, from the irregularity of $\mathcal{O}$ for $\mathcal{G}_0$ or $\mathcal{G}_0^-$ it follows that $\mathcal{O}$ is irregular for $\mathcal{G}_\rho^1$ and $\mathcal{G}_\rho^2$.*

Since bottom boundary points are always regular, the assertion of Corollary 2.3 easily follow from Lemmas 2.1 and 2.2 and from the fact that the regularity barrier for $\mathcal{O}$ regarded as a boundary point of $\mathcal{G}_\rho^2$ is at the same time a regularity barrier for $\mathcal{O}$ regarded as a boundary point of $\mathcal{G}_0^-$.

Obviously, the assertion of Corollary 2.3 is true if we take an arbitrary boundary point $z_0 = (x^0, t_0) \in \partial\mathcal{G}_0$, assuming that $\mathcal{G}_0^- = \{z \in \mathcal{G}_0 : t < t_0\} \ne \emptyset$. In this case we need to replace $\mathcal{G}_\rho^1$ and $\mathcal{G}_\rho^2$ with their translations after rigid body displacement composed of a translation in x-space and shift along the t-axis, in such a way that $\mathcal{O}$ coincides with z_0 after this displacement.

3. Proofs of the main results.

Proof of Theorem 1.1. The proof is based on the construction of the regularity barrier for $\mathcal{O}$ regarded as a boundary point of $\mathcal{G}^1_\rho$. We construct the regularity barrier assuming that $\delta > 0$ is sufficiently small. This produces no loss of generality, since boundary regularity is a local property. Without loss of generality we also assume that the positive numbers α and β satisfy

$$\alpha + 2(N-1)\beta = 1. \tag{3.1}$$

Indeed, if $\alpha + 2(N-1)\beta < 1$, then we can take $\tilde\beta > \beta$ such that $\alpha + 2(N-1)\tilde\beta = 1$ and consider the domain $\tilde{\mathcal{G}}^1_\rho$ by replacing β with $\tilde\beta$ in $\mathcal{G}^1_\rho$. It may be easily seen that $\tilde{\mathcal{G}}^1_\rho$ contains $\mathcal{G}^1_\rho$ if we replace δ in $\mathcal{G}^1_\rho$ with $\beta\tilde\beta^{-1}\delta$. Therefore, if we construct a regularity barrier for $\tilde{\mathcal{G}}^1_\rho$ for all $\delta \le \delta_1$, the latter will be a regularity barrier for $\mathcal{G}^1_\rho$ for all $\delta \le \beta\tilde\beta^{-1}\delta_1$.

Without loss of generality we may also assume that $\rho(\xi)$ is twice continuously differentiable for $\xi < 0$ and satisfies

$$\xi^2\rho^{-1}(\xi)\rho''(\xi) \to 0 \qquad \text{as } \xi \uparrow 0. \tag{3.2}$$

Indeed, otherwise we can choose a monotonically decreasing and twice continuously differentiable function $\rho_1(\xi), -\delta \le \xi < 0$, which satisfies the following conditions:

$$\frac{1}{2}\rho(\xi) < \rho_1(\xi) < \rho(\xi), \tag{3.3}$$

$$\min(\rho'(\xi); -\rho(\xi)) < \rho_1'(\xi) < \frac{1}{2}\rho'(\xi). \tag{3.4}$$

From (3.4) and (1.2) it follows that

$$0 < \xi\rho_1^{-1}(\xi)\rho_1'(\xi) < 2\max(\xi\rho^{-1}(\xi)\rho'(\xi); -\xi) \to 0 \qquad \text{as } \xi \uparrow 0. \tag{3.5}$$

Hence, from (3.3)–(3.5) it follows that ρ_1 satisfies (1.2) and (1.3). Applying l'Hôpital's rule from (3.5) and (3.4), we have

$$\frac{\xi\rho_1''(\xi)}{\rho_1'(\xi)} \to -1 \qquad \text{as } \xi \uparrow 0. \tag{3.6}$$

From (3.4)–(3.6) it easily follows that ρ_1 satisfies (3.2). Hence, ρ_1 satisfies all the required conditions, and in view of (3.3) we have $\mathcal{G}^1_\rho \subset \mathcal{G}^1_{\rho_1}$. Accordingly, the regularity barrier for $\mathcal{O}$ regarded as a boundary point of $\mathcal{G}^1_{\rho_1}$ will be a regularity barrier for $\mathcal{O}$ regarded as a boundary point of $\mathcal{G}^1_\rho$.

Thus it is enough to construct a regularity barrier for $\mathcal{O}$ regarded as a boundary point of $\mathcal{G}^1_\rho$, assuming additionally that ρ is C^2 for $\xi < 0$ and satisfies (3.2). By the way, all the examples of ρ from (1.4) satisfy these conditions as well.

We prove that the following function is the required regularity barrier:

$$u(x,t) = g(\xi)\exp\Big(-\frac{x_1^2}{4\xi}\Big) + \phi(\xi),$$

where ϕ is defined via

$$9\log\phi = \int_{\xi_0}^{\xi} \frac{\rho(\eta)}{\eta}d\eta, \qquad \xi_0 < \xi < 0,$$

with ξ_0 being a fixed negative number with sufficiently small $|\xi_0|$, and

$$g(\xi) = -\frac{1}{2}\rho(\xi)\phi(\xi).$$

From (1.2) and (1.3) it follows that

$$\phi(\xi) > 0, \quad 9\phi'(\xi) = \xi^{-1}\rho(\xi)\phi(\xi) < 0; \qquad \phi(\xi) \downarrow 0 \qquad \text{as } \xi \uparrow 0, \tag{3.7}$$

$$g(\xi) < 0, \quad g'(\xi) > 0; \qquad g(\xi) \uparrow 0 \qquad \text{as } \xi \uparrow 0. \tag{3.8}$$

The equation of the level hypersurface $u(x,t) = 0$ is given by

$$x_1^2 = 4\xi[\log\rho(\xi) - \log 2].$$

Moreover, we have

$$u > 0 \text{ in} \qquad \mathcal{G}' = \{z : x_1^2 < 4\xi[\log\rho(\xi) - \log 2], (\overline{x}, t) \in P(2\delta)\}.$$

Since $\mathcal{G}_\rho^1 \subset \mathcal{G}'$, we derive that u is positive and continuous in $\overline{\mathcal{G}_\rho^1}\backslash\{\mathcal{O}\}$. The function u is symmetric with respect to the x_1-variable, and for arbitrary fixed $(\overline{x}, t) \in \overline{P(\delta)}$, $(\overline{x}, t) \neq (0,0)$, u attains its maximum at $x_1 = 0$. Hence, we have

$$0 < u(x_1, \overline{x}, t) \le u(0, \overline{x}, t) = \phi(\xi)\Big(1 - \frac{1}{2}\rho(\xi)\Big) \to 0 \qquad \text{as } \xi \uparrow 0.$$

Thus u has a removable singularity at the point $\mathcal{O}$, and prescribing $u = 0$ at $\mathcal{O}$, we have $u \in C(\overline{\mathcal{G}_\rho^1})$. To complete the proof, we need to show that u is superparabolic in $\mathcal{G}_\rho^1$. Taking into account (3.1), we derive

$$\mathbf{D}u = \exp\Big(-\frac{x_1^2}{4\xi}\Big)\mathbf{S}, \tag{3.9}$$

where

$$\mathbf{S} = \frac{g(\xi)}{\xi}\Big[\frac{1}{2} + 2\beta^2|\overline{x}|^2 x_1^2\xi^{-2} - \frac{1}{4}\beta^2|\overline{x}|^2 x_1^4\xi^{-3}\Big] + g'(\xi)\Big[1 - 2\beta^2|\overline{x}|^2 x_1^2\xi^{-2}\Big]$$

$$+ \phi'(\xi)\exp\Big(\frac{x_1^2}{4\xi}\Big) - 4\beta^2|\overline{x}|^2\Big[g''(\xi) + \phi''(\xi)\exp\Big(\frac{x_1^2}{4\xi}\Big)\Big]. \tag{3.10}$$

Assuming that $|\xi|$ is sufficiently small, from (1.2) and (3.7) it follows that

$$\frac{g(\xi)}{\xi} - g'(\xi) > \frac{1}{2}\phi(\xi)\Big[\rho'(\xi) - \frac{1}{2}\frac{\rho(\xi)}{\xi}\Big] > 0. \tag{3.11}$$

Therefore, from (3.8), (3.10), and (3.11) we derive that

$$\mathbf{S} > \frac{1}{4}\frac{g(\xi)}{\xi} + \phi'(\xi)\exp\Big(\frac{x_1^2}{4\xi}\Big) + \frac{1}{4}\frac{g(\xi)}{\xi} - 4\beta^2|\overline{x}|^2\Big[g''(\xi) + \phi''(\xi)\exp\Big(\frac{x_1^2}{4\xi}\Big)\Big]. \tag{3.12}$$

Using (3.7), we estimate the first two terms on the right-hand side of (3.12) as follows:

$$\frac{1}{4}\frac{g(\xi)}{\xi} + \phi'(\xi)\exp\Big(\frac{x_1^2}{4\xi}\Big) \ge \frac{1}{4}\frac{g(\xi)}{\xi} + \phi'(\xi) = -\frac{\rho(\xi)\phi(\xi)}{72\xi} > 0.$$

Therefore, from (3.12) it follows that

$$\mathbf{S} > \frac{1}{4}\frac{g(\xi)}{\xi} - 4\beta^2|\overline{x}|^2\Big[g''(\xi) + \phi''(\xi)\exp\Big(\frac{x_1^2}{4\xi}\Big)\Big]. \tag{3.13}$$

Using (3.7) and (3.11) we easily derive that

$$9\phi'' = \frac{2}{\xi}\Big[\frac{g(\xi)}{\xi} - g'(\xi)\Big] < 0. \tag{3.14}$$

Since $\mathcal{G}_\rho^1 \subset \mathcal{G}'$, from (3.13), (3.14), and (3.7) we have

$$\mathbf{S} > \frac{1}{4}\frac{g(\xi)}{\xi} - 4\beta^2|\overline{x}|^2\Big[g''(\xi) + \frac{1}{2}\rho(\xi)\phi''(\xi)\Big] = \frac{1}{4}\frac{g(\xi)}{\xi}$$

$$+ 4\beta^2|\overline{x}|^2\Big[\frac{1}{2}\phi(\xi)\rho''(\xi) + \rho'(\xi)\phi'(\xi)\Big] > \frac{1}{4}\frac{g(\xi)}{\xi} + 2\beta^2|\overline{x}|^2\phi(\xi)\rho''(\xi). \tag{3.15}$$

If $|\overline{x}| = 0$, then from (3.15) and (3.8) it follows that $\mathbf{S} > 0$. Otherwise, from (3.15) we derive that

$$\mathbf{S} > 2\beta^2|\overline{x}|^2\phi(\xi)\Big[\frac{1}{16\beta}\frac{\rho(\xi)}{\xi^2} + \rho''(\xi)\Big]. \tag{3.16}$$

Assuming that $|\xi|$ is sufficiently small, from (3.2) and (3.16) we have $\mathbf{S} > 0$. Hence, from (3.9) it follows that u is superparabolic in $\mathcal{G}_\rho^1$ with sufficiently small δ. Accordingly, u is a required regularity barrier. The regularity of $\mathcal{O}$ for $\mathcal{G}_\rho^2$ is a consequence of Lemma 2.2. The theorem is proved.

The first assertion of Theorem 1.3 easily follows from Theorem 1.1 and Lemma 2.1. Indeed, since the space-time transformations due to rigid body displacements in the exterior $\mathcal{A}_\rho$-condition preserve the diffusion equation, the regularity barrier for $\mathcal{G}_\rho^2$ after the transformations according to exterior $\mathcal{A}_\rho$-condition will be a regularity barrier for $z_0 \in \mathcal{S}\Omega$. The second assertion of Theorem 1.3 is a consequence of the classical theory (see section 2).

4. Conclusions. The following natural question arises for each example of ρ from (1.4):

How sharp is the condition $\alpha + 2(N-1)\beta \le 1$ for the regularity of $\mathcal{O}$ for $\mathcal{G}_\rho^1$ or $\mathcal{G}_\rho^2$?

Let

$$\rho(\xi) = |\log|\xi||^{-1}. \tag{4.1}$$

In [P] it was proved that $\mathcal{O}$ is an irregular point for the bounded domain $\mathcal{G}_0$ lying in the strip $\{z : -\delta_1 < t < 0\}$ and bounded on its sides by the hypersurface of revolution

$$|x|^2 = -4\alpha_1 t\log_2|t|, \qquad \alpha_1 > 1.$$

It is easy to see that for sufficiently small δ_1, $\mathcal{G}_0 \subset \mathcal{G}_\rho^1$ if $1 < \alpha_1 < \alpha$. Therefore, from Corollary 2.3 it follows that $\mathcal{O}$ is irregular for $\mathcal{G}_\rho^1$ and $\mathcal{G}_\rho^2$ if $\alpha > 1, \beta > 0$. Obviously, the same result is true if we take any other particular example of ρ from (1.4).

Hence, the following natural question arises:

Is $\mathcal{O}$ regular or irregular for $\mathcal{G}_\rho^1$ and $\mathcal{G}_\rho^2$ if $\rho(\xi) = |\log|\xi||^{-1}$, $0 < \alpha \le 1$, $\beta > 0$, and $\alpha + 2(N-1)\beta > 1$?

The probabilistic analogue of this question was formulated at the end of section 1. This issue is addressed in a subsequent paper.

Another important conclusion says that $\mathcal{O}$ may be regular for

$$u_t = \Delta u \tag{4.2}$$

and at the same time irregular for

$$u_t = a\Delta u, \qquad 0 < a < 1, \tag{4.3}$$

regarded as a boundary point of $\mathcal{G}_\rho^1$ or $\mathcal{G}_\rho^2$. Let us check this fact by considering again the simplest case (4.1). Consider $\mathcal{G}_\rho^1$ and $\mathcal{G}_\rho^2$ with $\alpha = 1 - \frac{\varepsilon}{2}$, $\beta = \frac{\varepsilon}{4(N-1)}$, where ε is an arbitrary number satisfying $0 < \varepsilon \le 1 - a$. From Theorem 1.1 it follows that $\mathcal{O}$ is regular for (4.2) regarded as a boundary point of $\mathcal{G}_\rho^1$ or $\mathcal{G}_\rho^2$. If $u(x,t)$ solves (4.3) in $\mathcal{G}_\rho^1$, then after the transformation $x = x, \tau = at$, the function $\tilde{u}(x,\tau) = u(x,t)$ satisfies

$$\tilde{u}_\tau = \Delta \tilde{u}, \tag{4.4}$$

while the domain $\mathcal{G}_\rho^1$ is transformed to the domain

$$\tilde{\mathcal{G}}_\rho^1 = \{z = (x,\tau) : x_1^2 < 4\xi_1 \log \rho(\xi_1),\ -\delta < \xi_1 < 0,\ -\delta < \alpha a^{-1}\tau < 0\},$$

where $\xi_1 = \alpha a^{-1}\tau - \beta|\overline{x}|^2$. We have

$$\alpha a^{-1} > (1-\varepsilon)a^{-1} \ge 1.$$

Hence, $\mathcal{O}$ is irregular for (4.4) regarded as a boundary point of $\tilde{\mathcal{G}}_\rho^1$. Accordingly, $\mathcal{O}$ is irregular for (4.3) regarded as a boundary point of $\mathcal{G}_\rho^1$ or $\mathcal{G}_\rho^2$.

Acknowledgment. The author thanks Professor S. Luckhaus and Professor E. Zeidler for stimulating discussions.

REFERENCES

[B] H. Bauer, *Harmonische Räume und ihre Potentialtheorie*, Lecture Notes in Math. 22, Springer-Verlag, New York, 1966.

[D] J. L. Doob, *Classical Potential Theory and Its Probabilistic Counterpart*, Springer-Verlag, Berlin, 1984.

[EK] E. G. Effros and J. K. Kazdan, *On the Dirichlet problem for the heat equation*, Indiana Univ. Math. J., 20 (1971), pp. 683–693.

[EG] L. C. Evans and R. F. Gariepy, *Wiener's criterion for the heat equation*, Arch. Ration. Mech. Anal., 78 (1982), pp. 293–314.

[IM] K. Ito and H. P. McKean, *Diffusion Processes and Their Sample Paths*, Springer-Verlag, New York, 1996.

[L] E. M. Landis, *Necessary and sufficient conditions for regularity of a boundary point in the Dirichlet problem for the heat conduction equation*, Soviet Math., 10 (1969), pp. 380–384.

[P] I. G. Petrovsky, *Zur ersten Randwertaufgabe der Wärmeleitungsgleichung*, Compositio Math., 1 (1935), pp. 383–419.

[W] N. Wiener, *The Dirichlet problem*, J. Math. Phys., 3 (1924), pp. 127–146.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1435–1476

VORTEX MOTION LAW FOR THE SCHRÖDINGER–GINZBURG–LANDAU EQUATIONS*

DANIEL SPIRN†

Abstract. In the Ginzburg–Landau model for superconductivity a large Ginzburg–Landau parameter κ corresponds to the formation of tight, stable vortices. These vortices are located where an applied magnetic field pierces the superconducting bulk, and each vortex induces a quantized supercurrent about the vortex. The energy of large-κ solutions blows up near each vortex, which brings about difficulties in analysis. Rigorous asymptotic static theory has previously established the existence of a finite number of the vortices, and these vortices are located precisely at the critical points of a renormalized energy. We consider the motion of such vortices in a dynamic model for superconductivity that couples a $U(1)$ gauge-invariant Schrödinger-type Ginzburg–Landau equation to a Maxwell-type equation under the limit of large Ginzburg–Landau parameter κ. It is shown that under an almost-energy-minimizing condition each vortex moves in the direction of the net supercurrent located at the vortex position, and these vortices behave like point vortices in the classical two-dimensional Euler equations.

Key words. superconductivity, Ginzburg–Landau theory, vortex dynamics

AMS subject classifications. 35A20, 35B40, 35Q55, 82D55

PII. S0036141001396667

1. Introduction: Concentration phenomena and the Ginzburg–Landau equations.

Superconductivity has two important regimes, the Meissner state, where a superconductor completely repels a magnetic field, and the mixed-vortex, Shubnikov state, where the bulk is pierced by a magnetic field in small tubular regions called *vortices* or *filaments*. These normal-electron filaments are surrounded by large regions of superconducting Cooper-pairs, and each filament contains a quantized magnetic field inducing a circle of superconducting current around the filament. Both states of superconductivity can be effectively modeled with phase transition equations, called Ginzburg–Landau (GL) equations; see [15, 35].

Static GL equations were first proposed by V. L. Ginzburg and L. D. Landau [15] for a complex ordering parameter u and a magnetic field potential A. These equations allow for macroscopic deviations of the density $|u|^2$ of superconducting Cooper-pairs in the bulk, and the equations are derived from a free energy. Two length scales naturally arise out of the GL equations. The first is the *penetration depth* λ, which describes how far a magnetic field can penetrate the skin of the superconductor, and the second is the *coherence length* ξ, which measures the characteristic variation of the phase in the bulk. The important length scale is the *GL parameter* $\kappa = \lambda/\xi$. Although λ and ξ are temperature dependent, κ is mostly temperature independent and accurately describes the bifurcation between the Meissner state and the mixed-vortex state. After a suitable nondimensionalization the GL energy functional becomes

$$G(u,A) = \frac{1}{2}\int_\Omega |\nabla u - iAu|^2 + |\operatorname{curl} A - H_0|^2 + \frac{\kappa^2}{2}\left(1-|u|^2\right)^2 dx \tag{1.1}$$

*Received by the editors October 17, 2001; accepted for publication (in revised form) January 27, 2003; published electronically May 15, 2003.
http://www.siam.org/journals/sima/34-6/39666.html
†Department of Mathematics, Brown University, 151 Thayer St., Providence, RI 02912 (spirn@math.brown.edu).

and the corresponding GL equations are

$$0 = (\nabla - iA)^2 u + \kappa^2 u \left(1 - |u|^2\right),$$
$$0 = -\operatorname{curl} B + (iu, (\nabla - iA) u),$$

where $B = \operatorname{curl} A$ is the induced magnetic field. Here $(a, b) = \frac{1}{2}\left(a\bar{b} + \bar{a}b\right)$ denotes the complex inner product. When $\kappa < 1/\sqrt{2}$ then the superconductor is in the Meissner state, and when $\kappa > 1/\sqrt{2}$ the superconductor is in the mixed-vortex state. To capture coarse behavior of the GL equations with moderately large κ, as found in high T_C superconductors, it is important to understand the $\kappa \to \infty$ limit, and large κ solutions have tight, particle-like vortices as shown numerically in [23, 25].

1.1. Dynamic GL equations. We now introduce two separate dynamic models of superconductivity, each with a natural energy of the form (1.1). We will be interested in the first model.

Our model for a dynamic theory of superconductivity uses Schrödinger-type dynamics for the order parameter coupled to a Maxwell-type equation for the magnetic field potential. This Schrödinger–Ginzburg–Landau (SGL) model was first proposed in [29] based on arguments of Feynman [14]. The SGL equations retain gauge-invariance and can be viewed as a model for charged superfluids and other Bose–Einstein condensates which are coupled to Maxwell-type equations, such as in neutron stars. In addition to u and A there is an electric field potential Φ such that $E = \partial_t A + \nabla\Phi$ for the induced electric field E. The SGL system consists of [29]

$$\begin{aligned} \frac{1}{i}\left(\hbar\partial_t u + ie\Phi u\right) &= D\left(\hbar\nabla - ieA\right)^2 u + u\left(\beta|u|^2 + \alpha\right), \\ \beta^2\partial_t\left(\partial_t A + \nabla\Phi\right) + \delta\left(\partial_t A + \nabla\Phi\right) &= -\nu \operatorname{curl} B + 2\tau\left(iu, \left(\frac{\hbar}{2e}\nabla - iA\right)\right), \end{aligned} \tag{1.2}$$

where τ and D are microscopic parameters and ν^{-1} measures the conductivity of normal electrons. δ measures the normal conductivity of the medium, and superconducting alloys have δ of the order 10^{-3}, and β^2 measures relativistic effects and is of the order $10^{-9} \sim 10^{-11}$; see [29]. Since it would take an extremely long time to feel the effects of the β^2 term (far beyond the time frame of the asymptotics that follow), we set $\beta^2 = 0$. Suitably nondimensionalizing the SGL equations, we have

$$\begin{aligned} \frac{1}{i}\left(\partial_t u + i\Phi u\right) &= \left(\nabla - iA\right)^2 u + \kappa^2 u\left(1 - |u|^2\right), \\ \delta\left(\partial_t A + \nabla\Phi\right) &= -\operatorname{curl} B + (iu, (\nabla - iA) u) \end{aligned} \tag{1.3}$$

for δ small. The unusual coupling of a nonlinear Schrödinger equation to a parabolic equation for the magnetic field potential results in rather nontrivial behavior. When the electromagnetic field is not present, the equations become a nonlinear Schrödinger equation

$$\frac{1}{i}\partial_t u = \Delta u + \kappa^2 u\left(1 - |u|^2\right),$$

sometimes referred to as the Gross–Pitaevskii equation, especially in the context of the theory of superfluids.

A more widely studied dynamic model of superconductivity, called the time-dependent Ginzburg–Landau (TDGL) equations, can be formally derived from microscopic quantum theory [4, 5] and is sometimes referred to as the Gorkov–Eliashburg equations. The TDGL equations [29, 35] are

$$\begin{aligned}\hbar\partial_t u + ie\Phi u &= D\left(\hbar\nabla - ieA\right)^2 u + u\left(\beta|u|^2 + \alpha\right),\\ \partial_t A + \nabla\Phi &= -\nu \operatorname{curl} B + 2\tau\left(iu, \left(\frac{\hbar}{2e}\nabla - iA\right)u\right),\end{aligned}$$

where τ and D are microscopic parameters and ν^{-1} measures the conductivity of normal electrons. The TDGL equations are essentially a gradient flow of the GL functional (1.1) that preserves a *gauge* symmetry. After a suitable nondimensionalization we have the equations

$$\begin{aligned}\partial_t u + i\Phi u &= \left(\nabla - iA\right)^2 u + \kappa^2 u\left(1 - |u|^2\right),\\ \partial_t A + \nabla\Phi &= -\operatorname{curl} B + \left(iu, \left(\nabla - iA\right)u\right).\end{aligned}$$

We will use results on the asymptotic analysis of the TDGL equations [33] in section 3.

1.2. Vortices and quantization. We are interested in two-dimensional solutions to the SGL equations. A fundamental feature of GL model equations is the formation of vortices, which are locations where superconducting Cooper-pairs are locally absent, i.e., $|u|^2 = 0$. These normal-electron regions allow for quantized magnetic fields to penetrate through, and the quantized magnetic field induces a quantized supercurrent about it. Quantization phenomena have been observed experimentally; see [10, 11, 36]. Mathematically, a quantized supercurrent can be induced by defining the order parameter u over $\mathbb{C}$ and introducing a topological obstruction such that $|u| = 1$ and

$$\oint_{\partial B_r(x_0)} (iu, \partial_\tau u)\, d\omega = \deg\left(u, \partial B_r(x_0)\right) = d$$

with $d \in \mathbb{Z}$, $d \neq 0$. By a simple application of the Brouwer fixed point theorem there exists an $x \in B_r(x_0)$ such that $|u|^2(x) = 0$. Scaling arguments show that the radius of a vortex depends inversely on κ, and as κ becomes very large, a vortex can be described by the limiting point location and its winding number with an energy of order $\log\kappa$. We will be interested in understanding how these vortices move in the SGL equations, and we will look for simple equations of motion that describe the limiting point vortices as $\kappa \to \infty$. Such point vortex equations have been extensively studied in the context of the Euler equations; see [7].

1.3. Renormalized energy. As we will see, our vortex motion law relies on a quantity that depends on the vortex configuration and not the energy associated with an actual vortex. This free energy quantity $W(x_1, \ldots, x_d)$ is called the *renormalized energy* and is the total free energy minus the vortex self-induction energy. For the situation where there is no magnetic field potential A (see [1, 21]), we define the class

$$\mathcal{H}^1(x_j, \rho) = \begin{cases} u \in H^1\left(\Omega \setminus \bigcup_{j=1}^d B_\rho(x_j) : \mathbb{S}^1\right), \\ u = \frac{x - x_j}{|x - x_j|} \text{ on } \partial B_\rho(x_j) \text{ and } \partial_\nu u = 0 \text{ on } \partial\Omega, \end{cases}$$

then

$$W(\{x_j\}) = \lim_{\rho\to 0}\left\{\min_{u\in\mathcal{H}^1(x_j,\rho)}\frac{1}{2}\int_{\Omega\setminus\bigcup_{j=1}^d B_\rho(x_j)}|\nabla u|^2\,dx - \pi d\log\frac{1}{\rho}\right\} + \gamma d,$$

where γ is a universal constant described in [1]. This quantity depends solely on the location of the x_j's and Ω. A quick analysis of the renormalized energy shows that $W(x_j)\to\infty$ as $|x_k - x_l|\to 0$. A detailed analysis of the renormalized energy is performed in the appendix for the full GL energy functional (1.1) with Neumann boundary conditions.

1.4. Asymptotics and vortex motion laws. In the mathematical literature we replace κ^2 with $\frac{1}{\varepsilon^2}$ and study the $\varepsilon\to 0$ limit. It should be expected that the SGL equations would exhibit simple dynamic behavior in the $\varepsilon\to 0$ limit. Formal asymptotics [13, 28, 29] have reduced the equations to simple ODEs of the form

$$\frac{d}{dt}a_j = -\mathcal{J}\nabla_{a_j}\widetilde{W}(a_1,\dots,a_d), \tag{1.4}$$

where a_j is the location of the jth vortex and $\widetilde{W}(a_1,\dots,a_d)$ is a renormalized energy that depends on the precise asymptotic limit. Here

$$\mathcal{J} = \begin{pmatrix} 0 & -1\\ 1 & 0\end{pmatrix}.$$

So far, most efforts to make the $\varepsilon\to 0$ asymptotic limit of the SGL equations rigorous have revolved around the Gross–Pitaevskii equation of the form

$$\frac{1}{i}\partial_t u_\varepsilon = \Delta u_\varepsilon + \frac{1}{\varepsilon^2}u_\varepsilon\left(1-|u_\varepsilon|^2\right). \tag{1.5}$$

Equation (1.5) has been studied rigorously in the $\varepsilon\to 0$ limit in a series of papers [8, 24], where the vortex motion law (1.4) was rigorously derived for the limit of the simplified model equation (1.5) with both Neumann and Dirichlet boundary conditions.

Our aim is to study the $\varepsilon\to 0$ limit of the SGL equations with physically realistic boundary conditions and rigorously derive a vortex motion law. A crucial piece of information that strongly affects whether or not vortices follow this vortex motion law pertains to the precise amount of initial energy of the system. The SGL equations require a rigid bound, as excess energy will destroy the sensitive comparison arguments. This precise bound differs greatly from the TDGL equations, which require a much less stringent control on the initial data; see [33].

THEOREM 1.1. *Let $\{u_\varepsilon, A_\varepsilon\}$ solve the SGL equations*

$$\begin{aligned}\frac{1}{i}\left(\partial_t + i\Phi_\varepsilon\right)u_\varepsilon &= \left(\nabla - iA_\varepsilon\right)^2 u_\varepsilon + \frac{1}{\varepsilon^2}u_\varepsilon\left(1-|u_\varepsilon|^2\right),\\ \delta_\varepsilon\left(\partial_t A_\varepsilon + \nabla\Phi_\varepsilon\right) &= -\operatorname{curl} B_\varepsilon + \left(iu_\varepsilon, \left(\nabla - iA_\varepsilon\right)u_\varepsilon\right)\end{aligned}$$

under the Coulomb gauge in Ω such that $\partial_\nu u_\varepsilon = 0$, $\nu\cdot A_\varepsilon = 0$, and $B_\varepsilon = H_0$ on $\partial\Omega$. Let $\delta_\varepsilon\to 0$ and

$$\frac{\varepsilon^2|\log\varepsilon|}{\delta_\varepsilon}\to 0$$

as $\varepsilon \to 0$. *If the initial data* $\{u_\varepsilon(0), A_\varepsilon(0)\}$ *is chosen so that* d *vortices concentrate at* $\{a_k(0)\}$ *as* $\varepsilon \to 0$ *and satisfy the almost-energy-minimizing condition*

$$G_\varepsilon(u_\varepsilon, A_\varepsilon)(0) \le \pi d |\log \varepsilon| + W(\{a_k(0)\}) + o_\varepsilon(1),$$

then there will be d *vortices at* $\{a_k(t)\}$ *such that*

$$\frac{d}{dt} a_j = -\mathcal{J} \nabla_{a_j} W(\{a_k(t)\}),$$

where

$$\mathcal{J} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

and the renormalized energy $W(\{a_k\})$ *is defined by* (*A*.1).

1.5. Outline of paper. The following three sections of this paper can be loosely organized as variational theory, dynamic theory, and the proof of the vortex motion law.

In section 2 we review and generalize a few results on the GL energy functional. The SGL equations satisfy a set of conservation laws, shown in section 3, for the mass $|u_\varepsilon|^2$, the supercurrent or momentum j_ε, and the energy density g_ε. Since energy is slowly dissipative (and is conserved in the $\varepsilon \to 0$ limit), we are naturally led to use variational techniques for the GL energy functional (2.1). These techniques have been applied with great success to various dynamic GL equations; see [8, 19, 20, 24, 33]. We define structures, called essential zeros, where energy will concentrate. Outside of these essential zeros there is a uniform energy bound, and the scaled energy density will converge in measure to a sum of delta functions.

In section 3 we turn to the SGL equations and begin to study the dynamics. We start by showing that essential zeros move continuously in the $O(1)$ time scale. Outside of the essential zeros our uniform bound, along with the asymptotic scaling, yields the limiting expression for both the order parameter and the magnetic field potential. Next we establish a Γ-convergence result similar to results derived for the Gross–Pitaevskii equation; see [8, 24]. This convergence result controls the error of strong convergence by the amount of excess initial energy. At this point we have identified the asymptotic limit, modulo vortex position.

In section 4 we complete the vortex motion law proof by examining the conservation of momentum equation (3.11) in detail. We show that in the $\varepsilon \to 0$ limit

$$\frac{d}{dt} a_j(t) = -\mathcal{J} \nabla_{a_j} W(\{a_k(t)\}) + \nu,$$

where the defect measure ν results from the failure of strong convergence. We compare this limiting ODE to a solution of our desired ODE

$$\frac{d}{dt} b_j(t) = -\mathcal{J} \nabla_{b_j} W(\{b_k(t)\})$$

with $b_j(0) = a_j(0)$. If $\eta = \sum_{j=1}^d |a_j - b_j| \equiv 0$, then our task would be finished. To do so, we need to control the size of the defect measure, and this is accomplished with the Γ-convergence result, a careful restriction δ_ε of (1.3), and a Gronwall inequality.

In the appendix we derive the representations of the renormalized energy and its gradient. These proofs closely follow the methods of [2] for the renormalized

energy and [1] for the gradient of the renormalized energy. Finally, we compute the renormalized energy for a disk domain with d vortices.

Remark 1.2. A vortex motion law can also be rigorously derived for the Maxwell–Higgs equations, which are comprised of a gauge-invariant wave equation coupled to Maxwell's equations. It has the form

$$\begin{aligned}(\partial_t + i\Phi_\varepsilon)^2 u_\varepsilon &= (\nabla - iA_\varepsilon)^2 u_\varepsilon + \frac{1}{\varepsilon^2} u_\varepsilon \left(1 - |u_\varepsilon|^2\right),\\ \partial_t \left(\partial_t A_\varepsilon + \nabla \Phi_\varepsilon\right) &= -\operatorname{curl} B_\varepsilon + \left(iu_\varepsilon, (\nabla - iA_\varepsilon) u_\varepsilon\right).\end{aligned}$$

When time is rescaled $t \mapsto \sqrt{|\log \varepsilon|}t$, a motion law of the form

$$\frac{d^2}{dt^2} a_j = -\nabla_{a_j} W(\{a_k(t)\})$$

can be derived, and this result is communicated in [16].

2. Energy bounds and local regularity. We are interested in two-dimensional asymptotics; therefore $u : \mathbb{R}^2 \to \mathbb{C}$, $\Phi : \mathbb{R}^2 \to \mathbb{R}$, and $A : \mathbb{R}^2 \to \mathbb{R}^2$. We use interchangeably for $a \in \mathbb{R}$, $\operatorname{curl} a = (\partial_2 a, -\partial_1 a)$ and for $A \in \mathbb{R}^2$, $\operatorname{curl} A = \partial_1 A_2 - \partial_2 A_1$. Furthermore, we will take our domain $\Omega \subset \mathbb{R}^2$ to be a compact, simply connected set with smooth boundary. We define the covariant derivatives

$$\begin{aligned}\nabla_A &= \nabla - iA,\\ \partial_\Phi &= \partial_t + i\Phi\end{aligned}$$

for connections $\{A, \Phi\}$. These definitions will simplify notation and calculation. We will try to show as much as possible without fixing the gauge in an effort to increase the generality of our discussion.

2.1. $U(1)$ gauge. A solution $\{u, A, \Phi\}$ to a GL model equation is $U(1)$ gauge invariant if

$$\begin{aligned}u_\chi &= ue^{i\chi},\\ A_\chi &= A + \nabla\chi,\\ \Phi_\chi &= \Phi - \partial_t \chi\end{aligned}$$

is a solution to the same GL model equation. Although there is freedom to choose the gauge, physically relevant quantities are $U(1)$ gauge invariant. They are defined as follows.

DEFINITION 2.1. *We define the* mass $|u|^2$*, the* electric field $E = \partial_t A + \nabla\Phi$*, the* magnetic field $B = \operatorname{curl} A$*, the* supercurrent *or* momentum $j = j(u, A) = (iu, \nabla_A u)$*, the* charge $q = (iu, \partial_t u + iu\Phi)$*, and the* energy density

$$g_\varepsilon(u, A) = \frac{1}{2}\left[|\nabla_A u|^2 + |\operatorname{curl} A - H_0|^2 + \frac{1}{2\varepsilon^2}\left(1 - |u|^2\right)\right].$$

H_0 is a positive, finite constant.

We are given freedom to fix the gauge χ, and the various dynamic GL model equations are ill-posed if the gauge is not fixed; see [12]. In the dynamic setting of section 3 we will fix the *Coulomb gauge*, which makes $\operatorname{div} A = 0$ in Ω and $\nu \cdot A = 0$ on $\partial\Omega$.

PROPOSITION 2.2. *There exists a gauge choice such that* $\operatorname{div} A_\varepsilon = 0$ *for all* ε *with* $\nu \cdot A_\varepsilon = 0$ *on* $\partial\Omega$. *Furthermore,* $\partial_\nu u_\varepsilon = \partial_\nu \Phi_\varepsilon = 0$ *on* $\partial\Omega$.

Proof. Let χ be the $U(1)$ gauge, and let χ solve

$$\begin{aligned} \Delta\chi &= -\operatorname{div} A_\varepsilon \text{ in } \Omega, \\ \partial_\nu \chi &= -\nu \cdot A_\varepsilon \text{ on } \partial\Omega. \end{aligned}$$

Such a χ exists; see [2]. ∎

Unlike other gauges such as the temporal gauge ($\Phi \equiv 0$) or the Lorentz gauge (both $\partial_t \Phi + \operatorname{div} A = 0$ and $\Phi + \operatorname{div} A = 0$), the Coulomb gauge requires us to study and control the electric field potential Φ. However, the Coulomb gauge simplifies both the equation for the magnetic field potential and the representations of $W(\{a_k\})$ and $\nabla_{a_j} W(\{a_k\})$, both of which can be explicitly calculated for disk domains; see subsection A.3. See [12] for a more detailed discussion of the effect of gauge fixing on GL equations.

2.2. GL energy functional. In this subsection we examine the GL energy functional and establish various upper and lower bounds. Since we will be using energy comparison techniques, it is crucial to have estimates on energy minimizers. Energy minimizers $\{u_\varepsilon, A_\varepsilon\}$ of the GL energy functional

$$G_\varepsilon(u, A) = \frac{1}{2} \int_\Omega |\nabla_A u|^2 + |\operatorname{curl} A - H_0|^2 + \frac{1}{2\varepsilon^2} \left(1 - |u|^2\right)^2 dx \tag{2.1}$$

satisfy the static GL equations

$$\begin{aligned} 0 &= \nabla_{A_\varepsilon}^2 u_\varepsilon + \frac{1}{\varepsilon^2} u_\varepsilon \left(1 - |u_\varepsilon|^2\right)^2, \\ 0 &= -\operatorname{curl} B_\varepsilon + j_\varepsilon, \end{aligned}$$

which are the Euler–Lagrange equations of the energy functional. A rigorous treatment of the minimizing sequence with no electromagnetic field subject to Dirichlet boundary conditions was studied in [1], where the singular limit was completely characterized. The minimizing sequence with an applied magnetic field subject to Dirichlet boundary conditions was studied in [2], subject to Neumann boundary conditions with finite H_0 in [23] and with an asymptotically large H_ε in a series of papers [30, 31, 32]. We recall two important results on the full GL energy functional (2.1).

THEOREM 2.3 (Bethuel and Riviere [2]). *Let* $\{u_\varepsilon, A_\varepsilon\}$ *be a sequence of minimizers of the GL energy functional* (2.1) *on* $B_\rho(x_0)$ *such that* $|u_\varepsilon| = 1$ *on* $\partial B_\rho(x_0)$ *and*

$$\deg\left(u_\varepsilon, \partial B_\rho(x_0)\right) = \pm 1;$$

then

$$\pi |\log \varepsilon| - C(g, \rho) \le \int_{B_\rho(x_0)} g_\varepsilon(u_\varepsilon, A_\varepsilon) dx \le \pi |\log \varepsilon| + C(g, \rho).$$

Aside from this energy bound, [2] also characterized the limiting order parameter $u_\star$ and limiting magnetic field potential $A_\star$. $u_\star$ satisfies a harmonic map equation with a finite number of unit-valued positive vortices at $\{a_j\}_{j=1}^d$ equal to the winding number on the boundary, i.e., $d = \deg(u_\varepsilon, \partial\Omega)$. $A_\star$ satisfies a forced London equation $0 = \Delta B_\star - B_\star + 2\pi \sum_{j=1}^d \delta_{a_j}$, and the vortex positions minimize a renormalized energy;

see the appendix. A generalization of this result for any topologically constrained functions was established in [17].

THEOREM 2.4 (Jerrard). *Suppose $u_\varepsilon \in H^1_g(B_\rho, \mathbb{C})$ and $A_\varepsilon \in H^1(B_\rho, \mathbb{R}^2)$ such that $u_\varepsilon = g$ on ∂B_ρ. If $|g| = 1$, $\deg(g, \partial B_\rho) \neq 0$, and*

$$\int_{B_\rho} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \leq \pi |\log \varepsilon| + C,$$

then

$$\int_{B_\rho} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \geq \pi |\log \varepsilon| - C$$

and

$$\int_{B_\rho} B_\varepsilon^2 dx \leq C.$$

In order to study concentrations in dynamic GL equations, it is necessary to identify where a vortex will concentrate in the $\varepsilon \to 0$ limit, and to that end we define a structure that will ensure the formation of vortices.

DEFINITION 2.5. *A point $a_j \in \Omega$ is an* essential zero *for $\{u_\varepsilon, A_\varepsilon\}$ if there exists ε_0, α_j such that $\alpha_j \in [\alpha_0, 2\alpha_0]$ for $2\alpha_0 < 1$ and for all $\varepsilon < \varepsilon_0$*

$$\deg\left(\frac{u_\varepsilon}{|u_\varepsilon|}, \partial B_{\varepsilon^{\alpha_j}}(a_j)\right) = \pm 1, \tag{Es 1}$$

$$\varepsilon^{\alpha_j} \int_{\partial B_{\varepsilon^{\alpha_j}}(a_j)} g_\varepsilon(u_\varepsilon, A_\varepsilon)d\omega \leq \frac{\pi\left(d + \frac{1}{2}\right)}{\alpha_0}. \tag{Es 2}$$

Essential zeros were introduced in [34] and used in the context of TDGL equations in [20, 33] and the Gross–Pitaevskii equations in [24]. An essential zero is a natural tool, as it ensures the formation of a unit-valence vortex in the $\varepsilon \to 0$ limit and establishes the position of the vortex up to an ε^α error, in agreement with energy-minimizing sequences; see [9]. Furthermore, the structure theorem of [20] allows us to identify essential zeros with only energy bounds and control of $|\nabla u_\varepsilon|$. Note (Es 2) implies that on $\partial B_{\varepsilon^{\alpha_j}}(a_j)$ the degree is well-defined and $2 \geq |u_\varepsilon| \geq \frac{1}{2}$ for all ε small enough.

We now establish a global energy lower bound away from the essential zeros. Our aim is to identify the location of energy concentration with the location of the essential zeros. Set $\Omega_\varepsilon = \Omega \setminus \bigcup_{j=1}^d B_{\varepsilon^{\alpha_j}}(a_j)$.

LEMMA 2.6. *Let $\{u_\varepsilon, A_\varepsilon\}$, where $u_\varepsilon = \rho_\varepsilon e^{i\Theta_a + i\psi_\varepsilon}$, have essential zeros at $a_j \in \Omega$ with*

$$\min\left(\operatorname{dist}(a_j, a_k), \operatorname{dist}(a_j, \partial\Omega)\right) \geq \sigma > 0.$$

Let $G_\varepsilon(u_\varepsilon, A_\varepsilon) \leq \pi d |\log \varepsilon| + C_1$; then

$$\int_{\Omega_\varepsilon} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \geq \pi |\log \varepsilon| \sum_{j=1}^d \alpha_j - C_{12}, \tag{2.2}$$

$$\int_{\Omega_\varepsilon} |\nabla \rho_\varepsilon|^2 + \rho_\varepsilon^2 |\nabla \psi_\varepsilon - A_\varepsilon|^2 + |B_\varepsilon - H_0|^2 + \frac{1}{2\varepsilon^2}\left(1 - |u_\varepsilon|^2\right)^2 dx \leq C_{13}, \tag{2.3}$$

where C_{12} and C_{13} are functions of σ, Ω, and d.

Proof. Lemma 2.6 has been established in various forms, including [23, 16], and it is proved here. We prove the lower bound (2.2) by finding a lower bound for a minimizer, and to prove the lower bound for the minimizer, we first determine an upper bound for an energy minimizer $\{u, A\}$ on Ω_ε subject to the constraint that $\{u, A\} = \{u_\varepsilon, A_\varepsilon\}$ on each $\partial B_{\varepsilon^{\alpha_j}}(a_j)$.

1. Define the domains $\Omega_{2\varepsilon} = \Omega \backslash \bigcup_{j=1}^d B_{2\varepsilon^{\alpha_j}}(a_j)$ and $\Omega_{3\varepsilon} = \Omega \backslash \bigcup_{j=1}^d B_{3\varepsilon^{\alpha_j}}(a_j)$ and the comparison functions $\{u_{com}, A_{com}\}$, where

$$u_{com} = \begin{cases} u_{com}^1 & \text{in } \Omega_\varepsilon \backslash \Omega_{2\varepsilon}, \\ u_{com}^2 & \text{in } \Omega_{2\varepsilon} \backslash \Omega_{3\varepsilon}, \\ u_{com}^3 & \text{in } \Omega_{3\varepsilon}, \end{cases}$$

$$A_{com} = \begin{cases} A_{com}^1 & \text{in } \Omega_\varepsilon \backslash \Omega_{2\varepsilon}, \\ A_{com}^2 & \text{in } \Omega_{2\varepsilon} \backslash \Omega_{3\varepsilon}, \\ A_{com}^3 & \text{in } \Omega_{3\varepsilon}. \end{cases}$$

First set

$$u_{com}^3 = \prod_{j=1}^d \frac{x - a_j}{|x - a_j|} e^{i\overline{\psi}_\varepsilon^j} = e^{i\Theta_a} \prod_{j=1}^d e^{i\overline{\psi}_\varepsilon^j},$$
$$A_{com}^3 = \widetilde{A} = -\operatorname{curl} \widetilde{\xi}$$

on $\Omega_{3\varepsilon}$. Here $\overline{\psi}_\varepsilon^j$ is a constant to be set later, and $\widetilde{\xi}$ solves

$$-\Delta^2 \widetilde{\xi} + \Delta \widetilde{\xi} = 0$$

on Ω and $\widetilde{\xi} = 0$ and $\Delta \widetilde{\xi} = H_0$ on $\partial\Omega$. Direct calculation yields

$$\int_{\Omega_{3\varepsilon}} g_\varepsilon(u_{com}^3, A_{com}^3) dx \le \pi |\log \varepsilon| \sum_{j=1}^d \alpha_j + C_2(\Omega, \sigma, d). \tag{2.4}$$

2. We now want to show that $\int_{\Omega_{3\varepsilon} \backslash \Omega_\varepsilon} g_\varepsilon(u_{com}, A_{com}) dx \le C$. Without loss of generality we can compute the energy about one annulus centered at the origin. Start by defining $\{u_{com}^1, A_{com}^1\}$ as

$$A_{com}^1(r, \theta) = A_\varepsilon(\varepsilon^{\alpha_j}, \theta),$$
$$u_{com}^1(r, \theta) = \rho_{int} e^{i\Theta_a + i\psi_\varepsilon(\varepsilon^{\alpha_j}, \theta)},$$

where

$$\rho_{int}^2(r, \theta) = \left(\frac{2\varepsilon^{\alpha_j} - r}{\varepsilon^{\alpha_j}} \right) \rho_\varepsilon^2(\varepsilon^{\alpha_j}, \theta) + \left(\frac{r - \varepsilon^{\alpha_j}}{\varepsilon^{\alpha_j}} \right).$$

It follows that

$$\frac{1}{\varepsilon^2} \left(1 - \rho_{int}^2\right)^2 = \frac{1}{\varepsilon^2} \left(1 - \rho_\varepsilon^2\right)^2 \left(\frac{2\varepsilon^{\alpha_j} - r}{\varepsilon^{\alpha_j}} \right)^2 \le \frac{1}{\varepsilon^2} \left(1 - \rho_\varepsilon^2\right)^2 \tag{2.5}$$

for $r \in [\varepsilon^{\alpha_j}, 2\varepsilon^{\alpha_j}]$. Since (Es 2) implies $2 \geq \rho_\varepsilon \geq \frac{1}{2}$ for ε small enough, then

$$\begin{aligned} |\nabla \rho_{int}|^2 &= \frac{1}{4} \frac{\left|\nabla \rho_{int}^2\right|^2}{\rho_{int}^2} \\ &\leq \frac{1}{4} \frac{4\rho_\varepsilon^2 |\nabla \rho_\varepsilon|^2 + \varepsilon^{2-2\alpha_j} \frac{\left(1-\rho_\varepsilon^2\right)^2}{\varepsilon^2}}{\rho_\varepsilon^2 \left(\frac{2\varepsilon^{\alpha_j}-r}{\varepsilon^{\alpha_j}}\right) + \left(\frac{r-\varepsilon^{\alpha_j}}{\varepsilon^{\alpha_j}}\right)} \\ &\leq 4 \left[|\nabla \rho_\varepsilon|^2 + \varepsilon^{2-2\alpha_j} \frac{\left(1-\rho_\varepsilon^2\right)^2}{\varepsilon^2} \right]. \end{aligned} \tag{2.6}$$

Finally, since $1 \leq 4\rho_\varepsilon^2$ then

$$\rho_{int}^2 |\nabla \Theta_a + \nabla \psi_\varepsilon - A_\varepsilon|^2 \leq 4\rho_\varepsilon^2 |\nabla \Theta_a + \nabla \psi_\varepsilon - A_\varepsilon|^2. \tag{2.7}$$

Combining (2.5)–(2.7) yields

$$\begin{aligned} &\int_{B_{2\varepsilon^{\alpha_j}} \setminus B_{\varepsilon^{\alpha_j}}} g_\varepsilon(u_{com}^1, A_{com}^1) dx \\ &\quad \leq \int_{\varepsilon^{\alpha_j}}^{2\varepsilon^{\alpha_j}} \int_{\partial B_r} 4 \left[|\nabla \rho_\varepsilon|^2 + \varepsilon^{2-2\alpha_j} \frac{\left(1-\rho_\varepsilon^2\right)^2}{\varepsilon^2} \right] + 4\rho_\varepsilon^2 |\nabla \Theta_a + \nabla \psi_\varepsilon - A_\varepsilon|^2 \\ &\qquad + |B_\varepsilon - H_0|^2 + \frac{1}{2\varepsilon^2} \left(1-\rho_\varepsilon^2\right)^2 dw dr \\ &\quad \leq 4 \int_{\varepsilon^{\alpha_j}}^{2\varepsilon^{\alpha_j}} \int_{\partial B_r} g_\varepsilon(u_\varepsilon, A_\varepsilon) d\omega dr \\ &\quad \leq C_3(\Omega, \sigma, d). \end{aligned} \tag{2.8}$$

3. We now bound our comparison function in the next annulus. Define $\{u_{com}^2, A_{com}^2\}$ in $B_{3\varepsilon^{\alpha_j}} \setminus B_{2\varepsilon^{\alpha_j}}$ to be

$$\begin{aligned} A_{com}^2(r,\theta) = A_{int} &= \left(\frac{3\varepsilon^{\alpha_j} - r}{\varepsilon^{\alpha_j}}\right) A_\varepsilon(\varepsilon^{\alpha_j}, \theta) + \left(\frac{r - 2\varepsilon^{\alpha_j}}{\varepsilon^{\alpha_j}}\right) \widetilde{A}(3\varepsilon^{\alpha_j}, \theta), \\ u_{com}^2(r,\theta) &= e^{i\Theta_a + i\psi_{int}}, \end{aligned}$$

where

$$\psi_{int} = \left(\frac{3\varepsilon^{\alpha_j} - r}{\varepsilon^{\alpha_j}}\right) \psi_\varepsilon + \left(\frac{r - 2\varepsilon^{\alpha_j}}{\varepsilon^{\alpha_j}}\right) \overline{\psi}_\varepsilon^j$$

and $\overline{\psi}_\varepsilon^j = \fint_{\partial B_{\varepsilon^{\alpha_j}}} \psi_\varepsilon ds$. Then

$$\begin{aligned} |\nabla \Theta_a + \nabla \psi_{int} - A_{int}|^2 &\leq \left| \left(\frac{3\varepsilon^{\alpha_j} - r}{\varepsilon^{\alpha_j}}\right) (\nabla \Theta_a + \nabla \psi_\varepsilon - A_\varepsilon) + \left(\frac{r - 2\varepsilon^{\alpha_j}}{\varepsilon^{\alpha_j}}\right) \nabla \Theta_a \right. \\ &\qquad \left. + \nabla \left(\frac{r}{\varepsilon^{\alpha_j}}\right) \left(\psi_\varepsilon - \overline{\psi}_\varepsilon^j\right) - \left(\frac{r - 2\varepsilon^{\alpha_j}}{\varepsilon^{\alpha_j}}\right) \widetilde{A} \right|^2 \\ &\leq |\nabla \Theta_a + \nabla \psi_\varepsilon - A_\varepsilon|^2 + |\nabla \Theta_a|^2 + \varepsilon^{-2\alpha_j} \left|\psi_\varepsilon - \overline{\psi}_\varepsilon^j\right|^2 + \left|\widetilde{A}\right|^2 \\ &\leq 4 |\nabla_{A_\varepsilon} u_\varepsilon|^2 + |\nabla \Theta_a|^2 + \varepsilon^{-2\alpha_j} \left|\psi_\varepsilon - \overline{\psi}_\varepsilon^j\right|^2 + \left|\widetilde{A}\right|^2 \end{aligned}$$

and

$$|\operatorname{curl} A_{int} - H_0|^2 = \left|(B_\varepsilon - H_0)\left(\frac{3\varepsilon^{\alpha_j} - r}{\varepsilon^{\alpha_j}}\right) + \left(\frac{3\varepsilon^{\alpha_j} - r}{\varepsilon^{\alpha_j}}\right) H_0 + \left(\widetilde{A} - A_\varepsilon\right) \times \nabla\left(\frac{r}{\varepsilon^{\alpha_j}}\right) + \operatorname{curl} \widetilde{A}\left(\frac{r - 2\varepsilon^{\alpha_j}}{\varepsilon^{\alpha_j}}\right)\right|^2$$
$$\leq |B_\varepsilon - H_0|^2 + H_0^2 + \varepsilon^{-2\alpha_j} |A_\varepsilon|^2 + \varepsilon^{-2\alpha_j} \left|\widetilde{A}\right|^2 + \left|\operatorname{curl} \widetilde{A}\right|^2$$

for $r \in [2\varepsilon^{\alpha_j}, 3\varepsilon^{\alpha_j}]$. By Poincaré's inequality

$$\varepsilon^{-2\alpha_j} \int_{\partial B_{\varepsilon^{\alpha_j}}} |\psi_\varepsilon - \overline{\psi}_\varepsilon^j|^2 d\omega \leq C \int_{\partial B_{\varepsilon^{\alpha_j}}} |\nabla \psi_\varepsilon|^2 d\omega$$
$$\leq C \int_{\partial B_{\varepsilon^{\alpha_j}}} |\nabla \Theta_a + \nabla \psi_\varepsilon - A_\varepsilon|^2 + |\nabla \Theta_a| + |A_\varepsilon|^2 d\omega.$$

Then

$$\begin{aligned}
&\int_{B_{3\varepsilon^{\alpha_j}} \setminus B_{2\varepsilon^{\alpha_j}}} g_\varepsilon(u_{com}^2, A_{com}^2) dx \\
&\quad = \frac{1}{2} \int_{B_{3\varepsilon^{\alpha_j}} \setminus B_{2\varepsilon^{\alpha_j}}} |\nabla \Theta_a + \nabla \psi_{int} - A_{int}|^2 + |\operatorname{curl} A_{int} - H_0|^2 \, dx \\
&\quad \leq C \int_{\varepsilon^{\alpha_j}}^{2\varepsilon^{\alpha_j}} \int_{\partial B_r} |\nabla_{A_\varepsilon} u_\varepsilon|^2 + |B_\varepsilon - H_0|^2 + |\nabla \Theta_a|^2 \\
&\qquad + \varepsilon^{-2\alpha_j} \left[A_\varepsilon^2 + \widetilde{A}^2\right] + \left|\operatorname{curl} \widetilde{A}\right|^2 + H_0^2 \, d\omega dr \\
&\quad \leq C_4(\Omega, \sigma, d),
\end{aligned} \tag{2.9}$$

where we use the smoothness of $\widetilde{A}$ and Sobolev embedding for A_ε. Thus (2.4), (2.8), and (2.9) yield the upper bound

$$\int_{\Omega_\varepsilon} g_\varepsilon(u, A) dx \leq \int_{\Omega_\varepsilon} g_\varepsilon(u_{com}, A_{com}) dx \leq \pi |\log \varepsilon| \sum_{j=1}^{d} \alpha_j + C_5(\Omega, \sigma, d) \tag{2.10}$$

for our minimizer $\{u, A\}$. Note that our minimizer $\{u, A\}$ satisfies Neumann boundary conditions $\nu \cdot \nabla_A u = 0$ and $\operatorname{curl} A = H_0$ on $\partial\Omega$.

4. Next we show that a minimizer $\{u, A\}$ on $\overline{\Omega}_\varepsilon$ satisfies $|u| \geq \frac{1}{2}$. To do so we follow an argument in Lemma 2.2 of [21]. Suppose there exists $x_0 \in \Omega_\varepsilon$ such that $|u| < \frac{1}{2}$; then there are two cases. Suppose first that $\operatorname{dist}(x_0, \partial\Omega) \geq \varepsilon^{2\alpha_0}$; then $(B_{\varepsilon^{2\alpha_0}}(x_0) \cap \overline{\Omega}_\varepsilon) \cap \partial\Omega = \emptyset$. We choose $B_{\varepsilon^\beta}(x_0)$, where $\beta \in [2\alpha_0, 4\alpha_0]$ such that

$$\varepsilon^\beta \int_{\partial\left(B_{\varepsilon^\beta}(x_0) \cap \Omega_\varepsilon\right)} g_\varepsilon(u, A) d\omega \leq C$$

and such that $B_{\varepsilon^\beta}(x_0) \cap \Omega_\varepsilon$ does not intersect $\partial\Omega$. However, since $\{u, A\}$ is a minimizer with $\deg(u/|u|, \partial B_{\varepsilon^\beta}(x_0)) = 0$, then if there is a point x_0 such that $|u(x_0)| < 1/2$, then

$$\int_{B_{\varepsilon^\beta}(x_0)} g_\varepsilon(u, A) dx \geq (1 - \beta)\, \pi |\log \varepsilon| - C$$

for ε small enough; see [20, 21]. This contradicts (2.10) when α_0 is small.

If $\operatorname{dist}(x_0, \partial\Omega) \leq \varepsilon^{2\alpha_0}$, then choose $x_0^\star$ such that $|x_0 - x_0^\star| = \operatorname{dist}(x_0, \partial\Omega)$; then choose $\beta \in (\alpha_0, 2\alpha_0)$ such that $\varepsilon^\beta \int_{\partial B_{\varepsilon^\beta}(x_0^\star)} g_\varepsilon(u, A) dx \leq C$. Let $y = \Psi(x) = (\Psi_1(x), \Psi_2(x))$ be a conformal change of variables that maps $B_{\varepsilon^\beta}(x_0) \cap \Omega$ to B_1^+ such that $\partial\left(B_{\varepsilon^\beta}(x_0^\star) \cap \Omega\right) \cap B_{\varepsilon^\beta}(x_0^\star)$ is mapped to $\{y_2 = 0\}$. Then $u(x) = u'(\Psi(x)) = u'(y)$ and $A_k(x) = \left(\partial_{x_k} \Psi_j(x)\right) A_j'(\Psi(x)) = \left(\partial_{x_k} \Psi_j(x)\right) A_j'(y)$ satisfies

$$\begin{aligned} 0 &= \left(\nabla - iA'\right)^2 u' + \frac{c(y)}{\varepsilon^{2(1-\beta)}} u'(1 - |u'|^2), \\ 0 &= -\operatorname{curl} B' + c(y)\left(iu', \nabla u' - iA'u'\right) \end{aligned}$$

on B_1^+, where

$$c(y)^{-1} = \left(\left(\partial_{x_1}\Psi_1\right)^2 + \left(\partial_{x_1}\Psi_2\right)^2\right)$$

and $\partial_\nu u'(y) = 0$ on $\{y_2 = 0\}$. Our conformal map Ψ maps x_0 to a point y_0 in the interior of B_1^+. By constructing a suitable energy flow, we can show that $|u'| \to 0$ as $\varepsilon \to 0$ uniformly in B_1^+ by following [6], contradicting $|u'| < 1/2$ for all ε.

5. We can now complete the lower bound estimate for the minimizer $\{u, A\}$. Set $u = \rho e^{i(\Theta_a + \psi)}$; then by (2.10)

$$\pi |\log\varepsilon| \sum_{j=1}^d \alpha_j + C_5 \geq \frac{1}{2} \int_{\Omega_\varepsilon} |\nabla\rho|^2 + \rho^2 |\nabla\Theta_a + \nabla\psi - A|^2 + |B - H_0|^2 + \frac{1}{2\varepsilon^2}\left(1 - \rho^2\right)^2 dx.$$

We have the following crude estimates:

$$\begin{aligned} \frac{1}{2}\int_{\Omega_\varepsilon} \rho^2 |\nabla\Theta_a|^2 dx &= \frac{1}{2}\int_{\Omega_\varepsilon} \left(\rho^2 - 1\right)^2 |\nabla\Theta_a|^2 dx + \frac{1}{2}\int_{\Omega_\varepsilon} |\nabla\Theta_a|^2 dx \\ &\leq \left[\frac{|\Omega|}{4}\varepsilon\left(\varepsilon^{-2\alpha_0} + \frac{d-1}{\sigma^2}\right)\left(\int_{\Omega_\varepsilon}\left(1 - \rho^2\right)^2 dx\right)^{1/2}\right] \\ &\quad + \left[\pi |\log\varepsilon| \sum_{j=1}^d \alpha_j + C_6(\Omega, \sigma, d)\right] \\ &\leq \pi |\log\varepsilon| \sum_{j=1}^d \alpha_j + C_7(\Omega, \sigma, d) \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} &\int_{\Omega_\varepsilon} \left|1 - \rho^2\right| |\nabla\Theta_a| \, |\nabla\psi - A| \, dx \\ &\qquad \leq \frac{1}{8}\int_{\Omega_\varepsilon} |\nabla\psi - A|^2 dx + 2\int_{\Omega_\varepsilon} \left(1 - \rho^2\right)^2 |\nabla\Theta_a|^2 dx \\ &\qquad \leq \frac{1}{8}\int_{\Omega_\varepsilon} |\nabla\psi - A|^2 dx + 2C_7(\Omega, \sigma, d). \end{aligned} \tag{2.12}$$

Combining (2.11)–(2.12) yields

$$\begin{aligned} C_7 \geq &\int_{\Omega_\varepsilon} |\nabla\rho|^2 + \rho^2 |\nabla\psi - A|^2 + 2\nabla\Theta_a \cdot (\nabla\psi - A) \, dx \\ &+ \int_{\Omega_\varepsilon} |B - H_0|^2 + \frac{1}{2\varepsilon^2}\left(1 - \rho^2\right)^2 dx. \end{aligned} \tag{2.13}$$

To complete the lower bound it is necessary to control the term

$$\int_{\Omega_\varepsilon} \nabla\Theta_a \cdot (\nabla\psi - A)\, dx.$$

We fix the Coulomb gauge for the minimizer so that $\operatorname{div} A = 0$ in Ω and $\nu \cdot A = 0$ on $\partial\Omega$. Therefore, there exists a scalar function ξ such that $A = -\operatorname{curl}\xi$ in Ω and $\xi = 0$ on $\partial\Omega$. If $e^{i\Theta_a} = \prod_{j=1}^{d} e^{i\theta_{a_j}} = \prod_{j=1}^{d} \frac{x-a_j}{|x-a_j|}$ is the canonical harmonic map, then

$$\int_{\partial B_r(a_j)} \partial_\nu \Theta_a d\omega = \int_{\partial B_r(a_j)} \partial_\nu \theta_{a_j} d\omega = 0$$

and

$$\int_{\partial\Omega} \partial_\nu \Theta_a d\omega = -\sum_{j=1}^{d} \int_{\partial B_r(a_j)} \partial_\nu \theta_{a_j} d\omega = 0,$$

which, by the boundary conditions, implies

$$\int_{\partial\Omega} \partial_\nu \psi \, d\omega = 0.$$

Therefore,

$$\begin{aligned} \int_{\Omega_\varepsilon} \nabla\Theta_a \cdot (\nabla\psi + \operatorname{curl}\xi)\, dx &= \int_{\partial\Omega} \partial_\nu \Theta_a \left(\psi - \bar{\psi}\right) d\omega \\ &\quad + \sum_{j=1}^{d} \int_{\partial B_{\varepsilon^{\alpha_j}}(a_j)} \partial_\nu \Theta_a \left(\psi - \bar{\psi}_j\right) d\omega \\ &\quad + \sum_{j=1}^{d} \int_{\partial B_{\varepsilon^{\alpha_j}}(a_j)} \xi\, \partial_\tau \Theta_a d\omega \\ &\leq \frac{1}{8} \int_{\Omega_\varepsilon} |\nabla\psi|^2 + C_9 + C_{10}, \end{aligned} \tag{2.14}$$

where

$$\bar{\psi}_j = \fint_{B_{\varepsilon^{\alpha_j}}(a_j)} \psi \, d\omega$$

and

$$\bar{\psi} = \fint_{\partial\Omega} \psi \, d\omega,$$

which follows by using (Es 1) and (Es 2). Since $\rho > 1/2$ for the minimizer, then using (2.13) and (2.14) we find

$$\int_{\Omega_\varepsilon} |\nabla\psi + \operatorname{curl}\xi|^2 + |\Delta\xi - H_0|^2 \, dx \leq C_{11}(\Omega, \sigma, d),$$

which implies

$$\begin{aligned}\int_{\Omega_\varepsilon} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx &\geq \int_{\Omega_\varepsilon} g_\varepsilon(u, A)dx \\ &\geq \frac{1}{2}\int_{\Omega_\varepsilon} \rho^2 \left|\nabla \Theta_a\right|^2 dx - \frac{1}{2}\int_{\Omega_\varepsilon} \rho^2 \left|\nabla\psi + \operatorname{curl}\xi\right|^2 dx \\ &\geq \pi|\log\varepsilon| \sum_{j=1}^{d} \alpha_j - C_7 - C_{11},\end{aligned}$$

which proves (2.2).

6. To prove (2.3) we use an argument in the proof of Lemma 4.3 of [23], along with (2.2), (Es 1), and (Es 2), to show

$$\int_{B_{\varepsilon^{\alpha_j}}(a_j)} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \geq \pi|\log\varepsilon|\,(1-\alpha_j) + C$$

for each essential zero. Equation (2.3) follows directly. ∎

LEMMA 2.7. *Let $\{u_\varepsilon, A_\varepsilon\}$ have d essential zeros at a_j such that*

$$\min\left(\operatorname{dist}(a_j, a_k), \operatorname{dist}(a_j, \partial\Omega)\right) \geq \sigma > 0$$

and let $G_\varepsilon(u_\varepsilon, A_\varepsilon) \leq \pi d|\log\varepsilon| + C_1$; then for any $\varepsilon^{\alpha_j} < r < \frac{\sigma}{4}$

$$\pi\log\frac{r}{\varepsilon} - C_{14} \leq \int_{B_r(a_j)} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \leq \pi\log\frac{r}{\varepsilon} + C_{14},$$

where C_{14} depends on Ω, σ, and d.

Proof.

$$\begin{aligned}&\int_{B_r(a_j)} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \\ &\quad= \int_{B_r\setminus B_{\varepsilon^{\alpha_j}}(a_j)} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx + \int_{B_{\varepsilon^{\alpha_j}}(a_j)} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \\ &\quad\geq \frac{1}{2}\int_{B_r\setminus B_{\varepsilon^{\alpha_j}}(a_j)} \rho_\varepsilon^2 \left|\nabla\Theta_a\right|^2 dx - \frac{1}{2}\int_{B_r\setminus B_{\varepsilon^{\alpha_j}}(a_j)} \rho_\varepsilon^2 \left|\nabla\psi_\varepsilon - A_\varepsilon\right|^2 dx \\ &\qquad+ \int_{B_{\varepsilon^{\alpha_j}}(a_j)} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \\ &\quad\geq \frac{1}{2}\int_{B_r\setminus B_{\varepsilon^{\alpha_j}}(a_j)} \left|\nabla\Theta_a\right|^2 dx - C_7 - C_{13} + \pi|\log\varepsilon|\,(1-\alpha_j) - C_1 \\ &\quad= \pi|\log\varepsilon| - C_{14},\end{aligned}$$

where we used (2.11), Lemma 2.6, and Lemma 4.3 of [23]. The upper bound can be derived in the same way. ∎

If the essential zeros are well spaced, Lemma 2.7 yields a uniform global bound on the energy.

$$\int_{\Omega\setminus\bigcup_{j=1}^{d} B_r(a_j)} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \leq C_{14}(\Omega, \sigma, d, r). \tag{2.15}$$

Therefore, establishing the location of essential zeros affords global energy bounds away from vortex concentrations. We also find by a simple calculation that

$$\int_{\Omega\setminus\bigcup_{j=1}^{d} B_r(a_j)} |\nabla\psi_\star - A_\star|^2 + |B_\star - H_0|^2\, dx \le C_{15}(\Omega,\sigma,d,r), \tag{2.16}$$

where $\psi_\star$, $A_\star$, and $B_\star$ are the weak limits of ψ_ε, A_ε, and B_ε, respectively.

3. Asymptotic behavior of the SGL equations. We are interested in the vortex dynamics of the SGL equations under the Coulomb gauge. Many of the techniques in this section are based on the methods found in [8, 24]; however, there are a few difficulties with the SGL equations that do not appear in the Gross–Pitaevskii equation (1.5). The SGL equations are defined as

$$\begin{aligned} \frac{1}{i}\partial_{\Phi_\varepsilon} u_\varepsilon &= \nabla_{A_\varepsilon}^2 u_\varepsilon + \frac{1}{\varepsilon^2} u_\varepsilon \left(1 - |u_\varepsilon|^2\right), \\ \delta_\varepsilon E_\varepsilon &= -\operatorname{curl} B_\varepsilon + j_\varepsilon, \end{aligned}$$

where $E_\varepsilon = \partial_t A_\varepsilon + \nabla\Phi_\varepsilon$. We take $\delta_\varepsilon \to 0$ such that

$$\frac{\varepsilon^2 |\log\varepsilon|}{\delta_\varepsilon} \to 0. \tag{3.1}$$

We have natural boundary conditions

$$\begin{aligned} &\nu\cdot\nabla_{A_\varepsilon} u_\varepsilon = 0, \\ &B_\varepsilon = H_0, \\ &\nu\cdot E_\varepsilon = 0. \end{aligned}$$

Fixing the Coulomb gauge, the SGL equations become

$$\frac{1}{i}\partial_{\Phi_\varepsilon} u_\varepsilon = \nabla_{A_\varepsilon}^2 u_\varepsilon + \frac{1}{\varepsilon^2} u_\varepsilon \left(1 - |u_\varepsilon|^2\right), \tag{3.2}$$

$$\delta_\varepsilon E_\varepsilon = \Delta A_\varepsilon + j_\varepsilon \tag{3.3}$$

with new boundary conditions

$$\partial_\nu u_\varepsilon = 0, \tag{3.4}$$

$$B_\varepsilon = H_0, \tag{3.5}$$

$$\nu\cdot A_\varepsilon = 0, \tag{3.6}$$

$$\partial_\nu \Phi_\varepsilon = 0. \tag{3.7}$$

The Coulomb gauge simplifies our analysis of the magnetic field potential equation (3.3). We should note that global-in-time existence of $C(\mathbb{R}^+, H^2\otimes H^3)\cap C^1(\mathbb{R}^+, L^2\otimes H^1)$ solutions to (3.2)–(3.7) for a fixed ε can be shown by a long, but straightforward, modification of the proofs found in [3].

Since we are dealing extensively with covariant derivatives it is helpful to calculate the following commutator relationships.

PROPOSITION 3.1. *Covariant derivatives satisfy*

$$(\partial_\Phi\nabla_A - \nabla_A\partial_\Phi) = -iE, \tag{3.8}$$

$$\left(\nabla_A u, \nabla_A^2 u\right) = \operatorname{div}\left(\nabla_A u\otimes\nabla_A u\right) - \frac{1}{2}\nabla\left|\nabla_A u\right|^2 - B\left(j\times e_3\right), \tag{3.9}$$

where $\otimes$ is the usual fluid tensor

$$\operatorname{div}(v \otimes v) = \sum_{j=1} \partial_{x_j}(v_i, v_j)$$

for $v_k \in \mathbb{C}$.

Proof. The proof of (3.8) is a direct calculation. We now turn to (3.9).

$$\begin{aligned}
\left(\nabla_A u, \nabla_A^2 u\right) &= (\nabla_A u, \nabla \cdot \nabla_A u) - (\nabla_A u, iA \cdot \nabla_A u) \\
&= \operatorname{div}(\nabla_A u \otimes \nabla_A u) - (\nabla_A u, iA \cdot \nabla_A u) \\
&\quad - \sum_{j=1}^{2} (\partial_j \partial_k u - i \partial_j A_k u - i A_k \partial_j u, \partial_j u - i A_j u) \\
&= \operatorname{div}(\nabla_A u \otimes \nabla_A u) - \frac{1}{2} \nabla |\nabla_A u|^2 \\
&\quad - \sum_{j=1}^{2} (\partial_k A_j - \partial_j A_k)(iu, \partial_j u - i A_j u) \\
&\quad - \sum_{j=1}^{2} (i A_j \partial_k u - i A_k \partial_j u, \partial_j u - i A_j u) \\
&\quad - \sum_{j=1}^{2} \left(\partial_k u - i A_k u, i A_j \partial_j u + A_j^2 u\right) \\
&= \operatorname{div}(\nabla_A u \otimes \nabla_A u) - \frac{1}{2} \nabla |\nabla_A u|^2 - \operatorname{curl} A (j \times e_3). \qquad \square
\end{aligned}$$

3.1. Conservation laws. The SGL equations can be transformed into a series of conservation laws by using a variation of the Madelung transformation [26], used extensively in the study of the nonlinear Schrödinger equation. Recall the mass $|u|^2$, the momentum $j = (iu, \nabla_A u)$, and the energy density

$$g_\varepsilon(u, A) = \frac{1}{2} \left[|\nabla_A u|^2 + |\operatorname{curl} A - H_0|^2 + \frac{1}{2\varepsilon^2} \left(1 - |u|^2\right)^2 \right].$$

Then we have the following proposition.

PROPOSITION 3.2. *A solution $\{u_\varepsilon, A_\varepsilon, \Phi_\varepsilon\}$ to the SGL equations satisfies the conservation laws*

$$\frac{1}{2} \partial_t |u_\varepsilon|^2 = - \operatorname{div} j_\varepsilon = -\delta_\varepsilon \operatorname{div} E_\varepsilon, \tag{3.10}$$

$$\frac{1}{2} \partial_t j_\varepsilon = - \operatorname{div}(\nabla_{A_\varepsilon} u_\varepsilon \otimes \nabla_{A_\varepsilon} u_\varepsilon) + \nabla P_\varepsilon + \delta_\varepsilon B_\varepsilon (E_\varepsilon \times e_3) - \frac{1}{2} |u_\varepsilon|^2 E_\varepsilon, \tag{3.11}$$

$$\partial_t g_\varepsilon = -\delta_\varepsilon E_\varepsilon^2 + \operatorname{div}(\nabla_{A_\varepsilon} u_\varepsilon, \partial_{\Phi_\varepsilon} u_\varepsilon) + \operatorname{curl}(E_\varepsilon (B_\varepsilon - H_0)), \tag{3.12}$$

where

$$\operatorname{div}(\nabla_A u \otimes \nabla_A u) = \partial_j (\partial_k u - i A_k u, \partial_j u - i A_j u)$$

is the usual fluid tensor product and

$$P_\varepsilon = \frac{1}{2} \left[|\nabla_{A_\varepsilon} u_\varepsilon|^2 - B_\varepsilon^2 + \left(u_\varepsilon, \nabla_{A_\varepsilon}^2 u_\varepsilon\right) + \frac{1}{2\varepsilon^2} \left(1 - |u_\varepsilon|^4\right) \right] \tag{3.13}$$

is the pressure.

Proof. Mass conservation (3.10) and energy conservation (3.12) are direct calculations. We prove momentum conservation (3.11). By direct calculation we find

$$\partial_t j_\varepsilon = 2\left(i\partial_{\Phi_\varepsilon} u_\varepsilon, \nabla_{A_\varepsilon} u_\varepsilon\right) + \nabla q_\varepsilon - E_\varepsilon \left|u_\varepsilon\right|^2. \tag{3.14}$$

Plugging the SGL equations into (3.14) yields

$$\begin{aligned}
\partial_t j_\varepsilon &= -2\left(\nabla_{A_\varepsilon} u_\varepsilon, \nabla_{A_\varepsilon}^2 u_\varepsilon\right) + \frac{1}{2\varepsilon^2}\nabla\left(1-\left|u_\varepsilon\right|^2\right)^2 \\
&\quad + \nabla\left[\left(u_\varepsilon, \nabla_{A_\varepsilon}^2 u_\varepsilon\right) + \frac{1}{\varepsilon^2}\left|u_\varepsilon\right|^2\left(1-\left|u_\varepsilon\right|^2\right)\right] - E_\varepsilon\left|u_\varepsilon\right|^2 \\
&= -2\left[\operatorname{div}\left(\nabla_{A_\varepsilon} u_\varepsilon \otimes \nabla_{A_\varepsilon} u_\varepsilon\right) - \frac{1}{2}\nabla\left|\nabla_{A_\varepsilon} u_\varepsilon\right|^2 - B_\varepsilon\left(j_\varepsilon \times e_3\right)\right] \\
&\quad + \nabla\left(\frac{1}{2\varepsilon^2}\left(1-\left|u_\varepsilon\right|^4\right) + \left(u_\varepsilon, \nabla_{A_\varepsilon}^2 u_\varepsilon\right)\right) - E_\varepsilon\left|u_\varepsilon\right|^2,
\end{aligned}$$

where we used (3.9) in the first term. Finally we use

$$B_\varepsilon\left(j_\varepsilon \times e_3\right) = B_\varepsilon\left(\delta_\varepsilon E_\varepsilon + \operatorname{curl} B_\varepsilon\right) \times e_3$$

to complete the proof. □

Let $G_\varepsilon(u, A) = \int_\Omega g_\varepsilon(u, A)dx$; then (3.12) implies

$$G_\varepsilon(u_\varepsilon, A_\varepsilon)(t) + \delta_\varepsilon \int_0^t \int_\Omega E_\varepsilon^2 \, dxdt = G_\varepsilon(u_\varepsilon, A_\varepsilon)(0)$$

for all times. We require the initial data to satisfy the initial conditions

$$\left\{|u_\varepsilon(0)| \le 1/2\right\} \subseteq \bigcup_{j=1}^{d} B_{\delta_0}(a_j(0)) \subseteq \left\{\operatorname{dist}(x, \partial\Omega) \ge \delta_0\right\}, \tag{3.15}$$

$$G_\varepsilon(u_\varepsilon, A_\varepsilon)(0) \le \pi d|\log\varepsilon| + W(\{a_j(0)\}) + o_\varepsilon(1), \tag{3.16}$$

$$\{u_\varepsilon(0), A_\varepsilon(0)\} \text{ has } d \text{ essential zeros at } \{a_j(0)\} \tag{3.17}$$

so that

$$u_\varepsilon(0) \rightharpoonup \prod_{j=1}^{d}\left(\frac{x-a_j(0)}{|x-a_j(0)|}\right)^{d_j} e^{i\psi_0(x)}, \quad \text{where } d_j = \pm 1$$

weakly in $H^1_{loc}(\Omega\backslash\{a_1(0), \ldots, a_d(0)\})$ and $\psi_0 \in H^1(\Omega)$. These initial conditions are chosen to ensure the convergence of the initial data to the form (3.17) with vortices well spaced away from the boundary. The initial conditions (3.15)–(3.17) imply, for all $t > 0$,

$$G_\varepsilon(u_\varepsilon, A_\varepsilon)(t) \le G_\varepsilon(u_\varepsilon, A_\varepsilon)(0) \le \pi d|\log\varepsilon| + W(\{a_j(0)\}) + o_\varepsilon(1). \tag{3.18}$$

3.2. Energy concentration and weak compactness. We now wish to identify the weak limits of u_ε and A_ε, modulo vortex position, for any $t \in [0,T]$. To do so, we need to show that essential zeros move continuously in time. Then we use the asymptotic scaling, along with our uniform energy bounds, to identify the limiting functions.

LEMMA 3.3. *Suppose $\{u_\varepsilon, A_\varepsilon\}$ is a sequence of $H^1 \otimes H^1$ maps from $\Omega \subset \mathbb{R}^2$ into $\mathbb{C} \otimes \mathbb{R}^2$. Suppose*

$$G_\varepsilon(u_\varepsilon, A_\varepsilon) \leq \pi d |\log \varepsilon| + C$$

and suppose there are d essential zeros located at $\{a_j\}_1^d$ that satisfy

$$\min\{\operatorname{dist}(a_i, a_j), \operatorname{dist}(a_i, \partial\Omega)\} \geq \sigma > 0.$$

Then

$$\frac{g_\varepsilon(u_\varepsilon, A_\varepsilon)}{\pi |\log \varepsilon|} \rightharpoonup \sum_{j=1}^d \delta_{a_j}$$

in a Radon measure. Furthermore,

$$u_\varepsilon \rightharpoonup e^{i\Theta_a + i\psi_\star} = \prod_{j=1}^d \frac{x - a_j}{|x - a_j|} e^{i\psi_\star}$$

weakly in $H^1_{loc}(\Omega_a)$.

Proof. The proof follows from Lemma 2.7. ∎

We first establish the vortex motion law for the almost-energy-minimizing assumption. To do so, it will be necessary to trace the location of vortices as time progresses. Let $\Omega_a = \Omega \backslash \{a_j\}$; then we have the following.

PROPOSITION 3.4. *Under the assumptions, the linear momentum j_ε is uniformly bounded in $L^1_{loc}(\Omega)$ and up to subsequence*

$$j_\varepsilon \rightharpoonup v = \nabla\Theta_a + \nabla\psi_\star - A_\star$$

in $L^1_{loc}(\Omega_a)$, where $\operatorname{div} v = 0$ and $\Theta_a = \sum_{j=1}^d \operatorname{Arg} \frac{x-a_j}{|x-a_j|}$.

Proof. This follows directly from Lemma 3.3 and the gauge choice. ∎

LEMMA 3.5. *The linear momentum $j_\varepsilon \in L^1(\Omega)$ uniformly in ε. Let $\varphi \in C_0^\infty(\Omega)$, $\varphi = x_1$ for $x \in B_{R/2}(a_j)$, and $\varphi = 0$ for $x \notin B_R(a_j)$, where $R \in (0, \delta_0)$. Then for $a_j = (a_j^x, a_j^y)$,*

$$\int_{B_R(a_j)} \nabla^\perp \varphi \cdot j_\varepsilon dx \to 2\pi a_j^x.$$

Proof. We first note that $|u_\varepsilon| \in H^1(\Omega)$ uniformly in ε. Therefore, $|u_\varepsilon| \in L^p(\Omega)$ uniformly in ε for all $1 \leq p < \infty$. We wish to establish that $\nabla_{A_\varepsilon} u_\varepsilon \in L^q(\Omega)$ for $1 \leq q < 2$. Then $j_\varepsilon \in L^r(\Omega)$ for all $1 \leq r < 2$. Therefore, for $\phi \in C_0^\infty(\Omega)$

$$\begin{aligned}
\int_{B_R} \nabla^\perp \phi \cdot j_\varepsilon \, dx &\to \int_{B_R} \nabla^\perp \phi \cdot (\nabla\Theta_a + \nabla\psi_\star - A_\star) \, dx \\
&= \int_{B_R} \nabla^\perp \phi \cdot (\nabla\theta_j + A_\star) \, dx \\
&= \int_{B_\varepsilon(a_j)} \nabla^\perp \phi \cdot \nabla\theta_j dx + \int_{\partial B_R} x_1 \partial_\tau \theta_j dx - \int_{B_R} \nabla^\perp \phi \cdot A_\star dx.
\end{aligned}$$

To show L^p control of j_ε, we need to establish an energy bound. In order to do so we assume there is an essential zero at some point x_0. Then using arguments in Lemma 2.6 we can establish

$$\int_{B_r(x_0)} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \geq \pi \log \frac{r}{\varepsilon} - C.$$

Furthermore, a simple energy upper bound gives

$$\int_{B_{r'}(x_0)} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \leq \pi \log \frac{r'}{\varepsilon} - C.$$

Therefore, for $r > 2\varepsilon^\alpha$

$$\int_{B_{2r}(x_0)\setminus B_r(x_0)} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \leq C.$$

Then for $p \in [1, 2)$ we have

$$\begin{aligned}
\int_{B_1(x_0)} |\nabla_{A_\varepsilon} u_\varepsilon|^p \, dx &\leq \int_{B_{2\varepsilon^\alpha}} |\nabla_{A_\varepsilon} u_\varepsilon|^p \, dx + \sum_{j=1}^{d} \int_{B_{2^{j+1}\varepsilon^\alpha}\setminus B_{2^j\varepsilon^\alpha}} |\nabla_{A_\varepsilon} u_\varepsilon|^p \, dx \\
&\leq \left(2 \int_{B_{2\varepsilon^\alpha}} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx\right)^{p/2} \varepsilon^{(2-p)/2} \\
&\quad + \sum_{j=1}^{d} |B_{2^{j+1}\varepsilon^\alpha}\setminus B_{2^j\varepsilon^\alpha}|^{(2-p)/2} \\
&= o_\varepsilon(1) + C \sum_{j=1}^{d} \left(2^j \varepsilon^\alpha\right)^{2-p} \leq C. \qquad \square
\end{aligned}$$

We now wish to show the continuous motion of essential zeros in the SGL equations on the order 1 time scale.

LEMMA 3.6. *The essential zeros do not move on any slower time scale $t \sim o(1)$ as $\varepsilon \to 0$. On the time scale $t \sim O(1)$, the vortex locations $a_j^\varepsilon(t)$ are uniformly continuous in t as $\varepsilon \to 0$.*

Proof. We know

$$u_\varepsilon \rightharpoonup \prod_{j=1}^{d} \frac{x - a_j}{|x - a_j|} e^{i\psi_0(x)}$$

in $H^1_{loc}(\Omega_{a_0})$ with $\|\psi_0\|_{H^1(\Omega)} \leq C_0$. Let $R > 0$ be a small number, $R < \frac{\sigma}{4}$, where

$$\sigma = \min\{|a_l - a_j|, \operatorname{dist}(a_l, \partial\Omega), l, j = 1, \ldots, n, l \neq k\}.$$

If the essential zeros move continuously, then there exists $t_\star > 0$ such that for a fixed $R > 0$ small ($R < \frac{\sigma}{4}$) there exists an essential zero in each $B_{R/2}(a_k(0))$. Suppose the essential zeros are not continuous in time; then let λ_ε be the maximum time such that all essential zeros still lie in $B_{R/2}(a_j(0))$. Therefore, for some essential zero, $a_k(\lambda_\varepsilon) \notin B_{R/2}(a_k(0))$. Our aim is to show

$$\liminf_{\varepsilon \to 0^+} \lambda_\varepsilon > 0.$$

Rescale the time $t \mapsto \lambda_\varepsilon t$ and set

$$\begin{aligned}\widetilde{u}_\varepsilon(t) &= u_\varepsilon(\lambda_\varepsilon t),\\ \widetilde{A}_\varepsilon(t) &= A_\varepsilon(\lambda_\varepsilon t),\\ \widetilde{\Phi}_\varepsilon(t) &= \Phi_\varepsilon(\lambda_\varepsilon t).\end{aligned}$$

Then for all $t \in [0,1)$ the essential zeros lie within $B_{R/4}(a_j(0))$, so the SGL equations become

$$\begin{aligned}\frac{1}{i}\left(\partial_t \widetilde{u}_\varepsilon + i\lambda_\varepsilon \widetilde{\Phi}_\varepsilon \widetilde{u}_\varepsilon\right) &= \lambda_\varepsilon \nabla^2_{\widetilde{A}_\varepsilon}\widetilde{u}_\varepsilon + \frac{\lambda_\varepsilon}{\varepsilon^2}\widetilde{u}_\varepsilon\left(1-|\widetilde{u}_\varepsilon|^2\right),\\ \delta_\varepsilon\left(\partial_t \widetilde{A}_\varepsilon + \lambda_\varepsilon \nabla\widetilde{\Phi}_\varepsilon\right) &= \lambda_\varepsilon \Delta \widetilde{A}_\varepsilon + \lambda_\varepsilon \widetilde{j}_\varepsilon.\end{aligned}$$

Let $\widetilde{E}_\varepsilon = \partial_t \widetilde{A}_\varepsilon + \lambda_\varepsilon \nabla\widetilde{\Phi}_\varepsilon$ and $\partial_{\widetilde{\Phi}_\varepsilon}\widetilde{u}_\varepsilon = \partial_t\widetilde{u}_\varepsilon + i\lambda_\varepsilon\widetilde{\Phi}_\varepsilon\widetilde{u}_\varepsilon$ The momentum equation becomes

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\widetilde{j}_\varepsilon = &-\lambda_\varepsilon \operatorname{div}\left(\nabla_{\widetilde{A}_\varepsilon}\widetilde{u}_\varepsilon \otimes \nabla_{\widetilde{A}_\varepsilon}\widetilde{u}_\varepsilon\right) - \nabla\left(\lambda_\varepsilon \widetilde{P}_\varepsilon\right)\\ &+ \lambda_\varepsilon\delta_\varepsilon \widetilde{B}_\varepsilon\left(\widetilde{E}_\varepsilon \times e_3\right) - \frac{\lambda_\varepsilon}{2}|\widetilde{u}_\varepsilon|^2\widetilde{E}_\varepsilon\end{aligned} \tag{3.19}$$

and the energy equation becomes

$$\frac{d}{dt}g_\varepsilon(\widetilde{u}_\varepsilon,\widetilde{A}_\varepsilon) = -\frac{\delta_\varepsilon}{\lambda_\varepsilon}\widetilde{E}_\varepsilon^2 + \operatorname{div}\left(\nabla_{\widetilde{A}_\varepsilon}\widetilde{u}_\varepsilon, \partial_{\widetilde{\Phi}_\varepsilon}\widetilde{u}_\varepsilon\right) + \operatorname{curl}\left(\widetilde{E}_\varepsilon\left(\widetilde{B}_\varepsilon - H_0\right)\right). \tag{3.20}$$

Choose $\phi \in C_0^\infty(B_R(a_k(0)))$ such that

$$\phi^2 = \begin{cases} 1 & \text{in } B_{\frac{R}{2}-\delta}(a_k(0)),\\ 0 & \text{in } B_R \backslash B_{\frac{3R}{4}}(a_k(0)),\end{cases}$$

and linear in between, where δ is chosen so that each essential zero $a_j^\varepsilon \in B_{\frac{R}{2}-\delta}(a_j(0))$ for all $j \neq k$ for all $t \in [0,1]$. Multiplying equation (3.19) by $\nabla^\perp\phi$ and integrating over $B_{\frac{R}{2}}(a_k(0)) \times [0,1]$ yields

$$\begin{aligned}\int_{B_{\frac{R}{2}}(a_k(0))} \nabla^\perp\phi \cdot \widetilde{j}_\varepsilon\Big|_0^1 dx = &2\lambda_\varepsilon \int_0^1 \int_{B_{\frac{R}{2}}(a_k(0))}\left(\nabla_{\widetilde{A}_\varepsilon}\widetilde{u}_\varepsilon \otimes \nabla_{\widetilde{A}_\varepsilon}\widetilde{u}_\varepsilon\right) : \nabla\nabla^\perp\phi\, dxdt\\ &+ 2\lambda_\varepsilon \int_0^1 \int_{B_{\frac{R}{2}}(a_k(0))} \delta_\varepsilon \widetilde{B}_\varepsilon\left(\widetilde{E}_\varepsilon \times e_3\right) - \frac{|\widetilde{u}_\varepsilon|^2}{2}\widetilde{E}_\varepsilon\, dxdt.\end{aligned}$$

The left side converges to $a_k^x(1) - a_k^x(0)$ and the right side converges to 0. Likewise $a_k^y(1) - a_k^y(0) = 0$ implies $a_k(1) = a_k(0)$, and $\deg\left(\frac{\widetilde{u}_\varepsilon(1)}{|\widetilde{u}_\varepsilon(1)|}, \partial B_{\frac{R}{2}}(a_k(0))\right) = 1$. From (3.20) we get

$$\int_{B_{\frac{R}{2}}(a_k(0))} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \leq \pi|\log\varepsilon| + C_{14}.$$

By Proposition 3.7 below, we find a single essential zero within the ball $B_{\frac{R}{2}}(a_k(0))$, which contradicts our assumption. Therefore, $\liminf_{\varepsilon\to 0}\lambda_\varepsilon > 0$ implies that the essential zeros move continuously. ∎

To complete the proof of Lemma 3.6, we need to establish an essential zero with only an energy bound and control on the degree. There are a few results in this direction, and we point out [21], which establishes essential zeros for the simplified GL functional, and [17], which computes precise lower bounds for simplified and full GL functionals. We will follow the spirit of the former, which has less precise bounds but allows for identification of vortices up to order ε^α diameter.

PROPOSITION 3.7. *Set $B_R = B_R(x_0)$; then if $\{\widetilde{u}, \widetilde{A}\}$ satisfies $\int_{\overline{B}_R} g_\varepsilon(\widetilde{u}, \widetilde{A})dx \leq \pi|\log\varepsilon| + C$ and $\deg(\frac{\widetilde{u}}{|\widetilde{u}|}, \partial B_R) = 1$, there exists exactly one essential zero in B_R.*

Proof. In order to isolate the essential zero in B_R, we would like to use the structure theorem of [20]. A sufficient condition to use this theorem is the bound $|\nabla_{\widetilde{A}}\widetilde{u}|_{L^\infty(B_R)} \leq \frac{C}{\varepsilon}$, which we lack. To compensate we employ a short-time gradient flow, i.e., the TDGL equations

$$\begin{aligned} \partial_\Phi u &= \nabla_A^2 u + \frac{1}{\varepsilon^2} u\left(1 - |u|^2\right), \\ E &= -\operatorname{curl} B + (iu, \nabla_A u) \end{aligned}$$

in B_R with initial data $u(x,0) = \widetilde{u}$ if $|\widetilde{u}| \leq 1$ and $u(x,0) = \frac{\widetilde{u}}{|\widetilde{u}|}$ if $|\widetilde{u}| > 1$ and $A(x,0) = \widetilde{A}$, subject to boundary conditions

$$\begin{aligned} (iu, \nabla_A u)(x,t) &= \left(i\widetilde{u}, \nabla_{\widetilde{A}}\widetilde{u}\right), \\ \operatorname{curl} A(x,t) &= \widetilde{B} \end{aligned}$$

on ∂B_R for all $t \geq 0$. These boundary conditions are reasonable given the assumption $\int_{\partial B_R} g_\varepsilon(u, A)d\omega \leq C$. For a more detailed account of asymptotics of the TDGL equations, see [33]. The TDGL equations use a modified potential energy

$$\tilde{g}_\varepsilon(u, A) = \frac{1}{2}\left[|\nabla_A u|^2 + \left|\operatorname{curl} A - \widetilde{B}\right|^2 + \frac{1}{2\varepsilon^2}\left(1 - |u|^2\right)^2\right].$$

Since the gradient flow regularizes data sufficiently to use the structure theorem, we can find an essential zero at a short ε^2 time later. Furthermore, the TDGL equations pin essential zeros to their initial location for any time scale $o(|\log\varepsilon|)$ (see [20, 33]), which allows us to identify the essential zero at time $t = 0$. This method was used in Proposition 1.1 of [22] to study concentrations in the nonlinear wave equation. Let $t = \varepsilon^2$; then parabolic estimates imply $|\nabla_A u|_{L^\infty(B_R)}(\varepsilon^2) \leq \frac{C}{\varepsilon}$ (see Proposition 2.8 of [33]), and we can find an essential zero $a^\varepsilon \in B_R$ for $\left\{u(\varepsilon^2), A(\varepsilon^2)\right\}$ (see Claim 4.3 of [33]).

Next we show that at $t = 0$ an essential zero is located at the same point. From our parabolic energy bound $\int_{B_R} \widetilde{g}_\varepsilon(u, A)(t)dx \leq \pi|\log\varepsilon| + C$, we find

$$\begin{aligned} \int_{B_R}\left[\widetilde{g}_\varepsilon(u,A)(0) + \widetilde{g}_\varepsilon(u,A)(\varepsilon^2) + \int_0^{\varepsilon^2}|\partial_\Phi u|^2 + E^2 dt + \frac{1}{\varepsilon^2}\int_0^{\varepsilon^2}\widetilde{g}_\varepsilon(u,A)dt\right]dx \\ \leq 3\pi|\log\varepsilon| + C. \end{aligned}$$

Then at a^ε

$$\begin{aligned} \varepsilon^\alpha \int_{\partial B_{\varepsilon^\alpha}(a^\varepsilon)}\left[\tilde{g}_\varepsilon(u,A)(0) + \tilde{g}_\varepsilon(u,A)(\varepsilon^2) + \int_0^{\varepsilon^2}|\partial_\Phi u|^2 + E^2 dt + \frac{1}{\varepsilon^2}\int_0^{\varepsilon^2}\tilde{g}_\varepsilon(u,A)dt\right]d\omega \\ \leq C. \end{aligned}$$

We now rescale

$$\left\{\widehat{u}, \widehat{A}, \widehat{\Phi}\right\}(x,t) = \left\{u, \varepsilon^{\alpha} A, \varepsilon^2 \Phi\right\}(\varepsilon^{\alpha} x + a^{\varepsilon}, \varepsilon^2 t).$$

Then

$$\int_0^1 \int_{\partial B_1(0)} \widetilde{g}_{\tilde{\varepsilon}}(\widehat{u}, \widehat{A}) + \varepsilon^{2(\alpha-1)} \left|\partial_{\widehat{\Phi}} \widehat{u}\right|^2 + \varepsilon^{-2} \widehat{E}^2 d\omega dt \leq C, \tag{3.21}$$

where $\tilde{\varepsilon} = \varepsilon^{1-\alpha}$. Inequality (3.21) implies the degree is well-defined. Next, we get

$$\int_{\partial B_1(0)} \widetilde{g}_{\tilde{\varepsilon}}(\widehat{u}, \widehat{A})(0) + \widetilde{g}_{\tilde{\varepsilon}}(\widehat{u}, \widehat{A})(1) d\omega \leq C,$$

which implies

$$\deg\left(\frac{\widetilde{u}}{|\widetilde{u}|}, \partial B_{\varepsilon^{\alpha}}(a^{\varepsilon})\right) = \deg\left(\frac{\widehat{u}(0)}{|\widehat{u}(0)|}, \partial B_1(0)\right) = \deg\left(\frac{\widehat{u}(1)}{|\widehat{u}(1)|}, \partial B_1(0)\right).$$

Therefore, there is an essential zero. ∎

This implies, for all $t \in (0, t_\delta)$ such that $t_\delta = t_\delta(\sigma, d, \Omega)$,

$$\frac{g_\varepsilon(u_\varepsilon, A_\varepsilon)(t)}{\pi |\log \varepsilon|} \rightharpoonup \sum_{j=1}^{d} \delta_{a_j(t)}.$$

Since there are uniform bounds on the energy outside of the concentrations, we can identify the limiting $u_\star$ and $A_\star$.

PROPOSITION 3.8. *The function $\psi_\star(x,t)$ satisfies in the weak limit of Proposition* 3.4

$$\Delta \psi_\star = 0$$

in Ω and $\partial_\nu \psi_\star = -\partial_\nu \Theta_a$ on $\partial\Omega$.

Proof. First, we let $\phi(x) \in C_0^\infty(\Omega)$ and $\varphi(t) \in C_0^\infty(0,T)$; then

$$\lim_{\varepsilon \to 0} \int_0^T \varphi(t) dt \int_{\Omega_a} j_\varepsilon \, \phi(x) dx = \int_0^T \varphi(t) dt \int_{\Omega_a} \left(\nabla \Theta_a + \nabla \psi_\star - A_\star\right) \phi dx, \tag{3.22}$$

and using (3.10)

$$\begin{aligned} \int_0^T \varphi(t) dt \int_{\Omega_a} j_\varepsilon \cdot \nabla \phi(x) dx &= \int_0^T \partial_t \varphi(t) dt \int_{\Omega_a} \frac{|u_\varepsilon|^2}{2} dx \\ &\to 0. \end{aligned}$$

Therefore, j_ε's weak limit is divergence free for a.e. $t \in [0,T]$. Combining with (3.22) implies

$$\Delta \psi_\star = 0$$

in Ω_a since $\operatorname{div} A_\varepsilon = 0$ for all ε. From Lemma 2.6 we find that $\psi_\star$ has removable singularities at the vortices. Therefore, $\psi_\star$ is harmonic throughout Ω.

To establish the Neumann boundary conditions we rely on a method of [24]. Near the boundary $\partial\Omega$ there are no essential zeros by Lemma 3.6. Choose $\phi \in C^\infty(B_r(x))$ such that $B_r(x) \cap \partial\Omega \neq \emptyset$ and $r \leq \frac{\sigma}{4}$. Since $u_\varepsilon = \rho_\varepsilon e^{i(\Theta_a+\psi_\varepsilon)}$, then the boundary conditions under the Coulomb gauge reduce to $\nu\cdot(\nabla\rho_\varepsilon + i\rho_\varepsilon\nabla(\Theta_a+\psi_\varepsilon))e^{i(\Theta_a+\psi_\varepsilon)} = 0$, $\nu\cdot A_\varepsilon = 0$, and $B_\varepsilon = H_0$. This implies, for each $\varepsilon > 0$,

$$\partial_\nu\rho_\varepsilon = \partial_\nu(\Theta_a + \psi_\varepsilon)) = 0$$

for $x \in \partial\Omega$. Now using $\operatorname{div} j_\varepsilon = 0$ on $x \in \partial\Omega$, then

$$\begin{aligned}\int_{\partial\Omega} \phi\,\nu\cdot v\,d\omega &= \int_\Omega v\cdot\nabla\phi\,dx = \lim_{\varepsilon\to 0}\int_\Omega j_\varepsilon\cdot\nabla\phi\,dx \\ &= \lim_{\varepsilon\to 0}\int_{\partial\Omega}\phi\,\nu\cdot j_\varepsilon\,d\omega + \lim_{\varepsilon\to 0}\int_\Omega \phi\,\partial_t\frac{|u_\varepsilon|^2}{2}dx,\end{aligned}$$

where $v = \nabla\Theta_a + \nabla\psi_\star - A_\star$. Integrating over $[0,t]$ yields

$$\int_0^t\int_{\partial\Omega}\phi\,\nu\cdot v\,dx = \lim_{\varepsilon\to 0}\frac{1}{2}\int_0^t\int_{\Omega_a}|u_\varepsilon|^2\partial_t\phi\,dx = 0.$$

Then we get $v\cdot\nu = 0$ on $\partial\Omega$. □

PROPOSITION 3.9. *Let $\{u_\varepsilon, A_\varepsilon\}$ be a solution to the SGL equations with scaling $\delta_\varepsilon \to 0$; then $A_\star = -\operatorname{curl}\xi_\star$ satisfies in the weak limit of Proposition* 3.4

$$\Delta^2\xi_\star - \Delta\xi_\star + 2\pi\sum_{j=1}^d \delta_{a_j(t)} = 0 \tag{3.23}$$

in Ω and $\xi = 0$ and $\Delta\xi = H_0$ on $\partial\Omega$.

Proof. From the lower bound Lemma 2.6 and the initial condition, (3.18) implies

$$\delta_\varepsilon\int_0^T\int_\Omega E_\varepsilon^2\,dxdt \leq C. \tag{3.24}$$

Therefore, we find that

$$\int_0^T\int_\Omega |-\operatorname{curl} B_\varepsilon + j_\varepsilon|^2\,dxdt \to 0$$

in $L^2(\Omega)$. Multiply (3.3) by $\nabla^\perp\phi$, where $\phi \in C_0^\infty(\Omega_a)$; then

$$\Delta B_\star - B_\star = 0$$

in distribution. Next, let $\phi \in C_0^\infty(B_\delta(a_j))$ for some a_j, and

$$\begin{aligned}0 &= \int_{B_\delta(a_j)}\nabla^\perp\phi\cdot E_\varepsilon + \int_{B_\delta(a_j)}\nabla^\perp\phi\,(-\operatorname{curl} B_\varepsilon + j_\varepsilon)\,dx \\ &\to \int_{B_\delta(a_j)}\phi\,(\Delta B_\star - B_\star)\,dx + 2\pi\phi(a_j),\end{aligned}$$

and so

$$\Delta B_\star - B_\star + 2\pi\delta_{a_j(t)} = 0$$

in distribution. Therefore, $B_\varepsilon \rightharpoonup B_\star$ in $H^1_{loc}(\Omega_a)$ and $B_\star$ satisfies

$$\Delta B_\star - B_\star + 2\pi \sum_{j=1}^{d} \delta_{a_j(t)} = 0$$

in distribution.

We now establish the boundary value of $B_\star$. Let $\phi = (\phi_1, \phi_2)$, where $\phi_j \in C^\infty(B_r(x))$, where $B_r \cap \partial\Omega \neq \emptyset$ and $r \leq \frac{\sigma}{4}$. Therefore, ϕ intersects part of the boundary and is supported away from any essential zero. Therefore,

$$\begin{aligned}
\int_{\partial\Omega} (B_\star - H_0)\, \phi \cdot \tau ds &= \int_{\partial\Omega} (B_\star - H_0)\, \phi \cdot dl \\
&= \int_\Omega \operatorname{curl} (\phi\, (B_\star - H_0))\, dx \\
&= \int_\Omega (B_\star - H_0) \operatorname{curl} \phi dx + \int_\Omega \phi \cdot \operatorname{curl} B_\star dx \\
&= \int_\Omega (B_\star - H_0) \operatorname{curl} \phi dx + \int_\Omega \phi \cdot j_\star dx \\
&= \lim_{\varepsilon\to 0} \int_\Omega (B_\varepsilon - H_0) \operatorname{curl} \phi + \phi \cdot j_\varepsilon dx \\
&= \lim_{\varepsilon\to 0} \int_\Omega \operatorname{curl} (\phi\, (B_\varepsilon - H_0)) + \delta_\varepsilon\, \phi \cdot E_\varepsilon dx \\
&= \lim_{\varepsilon\to 0} \int_{\partial\Omega} (B_\varepsilon - H_0)\, \phi \cdot dl + \lim_{\varepsilon\to 0} \delta_\varepsilon \int_\Omega \phi \cdot E_\varepsilon dx \\
&= 0
\end{aligned}$$

for a.e. $t \in [0, T]$. ∎

3.3. Γ-convergence. Unlike the TDGL equations, to establish strong convergence of the SGL equations we need a Γ-convergence-type result in the spirit of [8] and [24]. This result will help us twofold. First, the Γ-convergence will ensure strong convergence along the chosen subsequence, away from essential zeros, to the canonical harmonic map. Second, the Γ-convergence will be used to close a Gronwall inequality, critical in the proof of the vortex motion law.

LEMMA 3.10. *Let $\{u_\varepsilon, A_\varepsilon\}$ have essential zeros at vortex locations $\{a_1, \dots, a_d\}$. If, for some $\mu > 0$,*

$$\limsup_{\varepsilon\to 0} [G_\varepsilon(u_\varepsilon, A_\varepsilon) - \pi d |\log \varepsilon|] \leq \pi W(\{a_j\}) + \mu,$$

then for any $r > 0$, there is a constant C independent of ε and r such that for any $t > 0$

$$\limsup_{\varepsilon\to 0} \left\| \frac{j_\varepsilon}{|u_\varepsilon|} - v \right\|^2_{L^2(\Omega_r)} \leq C\mu, \tag{3.25}$$

$$\limsup_{\varepsilon\to 0} \|B_\varepsilon - B_a\|^2_{L^2(\Omega_r)} \leq C\mu, \tag{3.26}$$

$$\limsup_{\varepsilon\to 0} \|\nabla |u_\varepsilon|\|^2_{L^2(\Omega_r)} \leq C\mu, \tag{3.27}$$

where $\Omega_r = \Omega \backslash \bigcup_{j=1}^d B_r(a_j)$. *Here* $v = \nabla\Theta_a + \nabla\psi_a - A_a$ *such that* $\Delta\psi_a = 0$ *in* Ω, $\partial_\nu\psi_a = -\partial_\nu\Theta_a$ *on* $\partial\Omega$. *Furthermore,* $B_a = \Delta\xi_a$ *and* $A_a = -\operatorname{curl}\xi_a$ *such that* $\Delta^2\xi_a - \Delta\xi_a + 2\pi\sum_{j=1}^d \delta_{a_j} = 0$ *in* Ω, $\xi_a = 0$, $\Delta\xi_a = H_0$ *on* $\partial\Omega$.

Proof. The idea is to cut out small balls of radius ρ that contain d essential zeros. Inside of each of these balls, we replace $\{u_\varepsilon, A_\varepsilon\}$ with a minimizer on a slightly smaller ball subject to canonical boundary conditions on the boundary of the smaller ball and a simple interpolation in the annulus between the two balls. We can then use knowledge of energy minimizers from [2] and the renormalized energy of the appendix to control the strong convergence error.

1. By Lemma 2.6 we have that $u_\varepsilon \rightharpoonup e^{i(\Theta_a+\psi_\star)} = \prod_{j=1}^d \frac{x-a_j}{|x-a_j|} e^{i\psi_\star}$ weakly in $H^1(\Omega_a)$ for some $\nabla\psi_\star \in L^2_{loc}(\Omega_a)$, $A_\varepsilon \rightharpoonup A_\star$ weakly in $H^1(\Omega)$, and $B_\varepsilon \rightharpoonup B_\star$ weakly in $L^2(\Omega)$. Therefore,

$$\frac{j_\varepsilon}{|u_\varepsilon|} \rightharpoonup \nabla\Theta_a + \nabla\psi_\star - A_\star$$

weakly in $L^2(\Omega_a)$, and for any $\rho > 0$ and all $\varepsilon \le \varepsilon_0(\rho)$ small enough

(3.28)
$$\begin{aligned}\int_{\Omega_\rho} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx &= \frac{1}{2}\int_{\Omega_\rho}\left[|\nabla|u_\varepsilon||^2 + \left|\frac{j_\varepsilon}{|u_\varepsilon|}\right|^2 + |B_\varepsilon - H_0|^2 + \frac{1}{2\varepsilon^2}\left(1-|u_\varepsilon|^2\right)^2\right]dx \\ &\ge \frac{1}{2}\int_{\Omega_\rho} |\nabla|u_\varepsilon||^2 + \left|\frac{j_\varepsilon}{|u_\varepsilon|} - (\nabla\Theta_a + \nabla\psi_\star - A_\star)\right|^2 + |B_\varepsilon - B_\star|^2\,dx \\ &\quad + \frac{1}{2}\int_{\Omega_\rho} |\nabla\Theta_a + \nabla\psi_\star - A_\star|^2\,dx + |B_\star - H_0|^2 + o_\varepsilon(1),\end{aligned}$$

where $\Omega_\rho = \Omega \backslash \bigcup_{j=1}^d B_\rho(a_j)$.

2. We now choose $\rho \in (\frac{r}{2}, r)$ such that $\int_{\partial B_\rho} g_\varepsilon(u_\varepsilon, A_\varepsilon)d\omega \le C$; then

$$\begin{aligned} u_\varepsilon &\rightharpoonup e^{i\Theta_a + i\psi_\star}, \\ A_\varepsilon &\rightharpoonup A_\star \end{aligned}$$

weakly in $H^1(\partial B_\rho(a_j))$. We want to show that the $\{u_\varepsilon, A_\varepsilon\}$ is close to a canonical harmonic map inside of $B_\rho(a_j)$. In particular we want to define a comparison map $\{u_\varepsilon^\rho, A_\varepsilon^\rho\}$ that satisfies

$$\int_{B_\rho(a_j)} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx \ge \int_{B_\rho(a_j)} g_\varepsilon(u_\varepsilon^\rho, A_\varepsilon^\rho)dx + o(\rho,\varepsilon), \tag{3.29}$$

$$\int_{B_\rho(a_j)} g_\varepsilon(u_\varepsilon^\rho, A_\varepsilon^\rho)dx \ge \log\frac{\rho}{\varepsilon} + \gamma + o(\rho,\varepsilon). \tag{3.30}$$

We let

$$\{u_\varepsilon^\rho, A_\varepsilon^\rho\} = \begin{cases} \{u_\varepsilon, A_\varepsilon\} & \text{on } \Omega_\rho = \Omega\backslash\bigcup_{j=1}^d B_\rho(a_j), \\ \{u_{int}, A_{int}\} & \text{on each } B_\rho\backslash B_{\tilde\rho}, \\ \{u_{min}, A_{min}\} & \text{on each } B_{\tilde\rho} \end{cases}$$

such that

$$\{u_{min}, A_{min}\} = \min\left\{(u,A) \in H^1_{\psi_j}\left(B_{\tilde\rho}(a_j)\right) \otimes H^1_{\overline{B}_j}\left(B_{\tilde\rho}(a_j)\right)\right\},$$

where

$$H^1_{\overline{\psi}_j}(B_{\widetilde{\rho}}(a_j)) = \left\{ u \in H^1 : B_{\widetilde{\rho}} \to \mathbb{C} \text{ such that } u = \frac{x-a_j}{|x-a_j|} e^{i\overline{\psi}_j} \text{ on } \partial B_{\widetilde{\rho}}(a_j) \right\},$$
$$H^1_{\overline{B}_j}(B_{\widetilde{\rho}}(a_j)) = \left\{ A \in H^1 : B_{\widetilde{\rho}} \to \mathbb{R}^2 \text{ such that } \operatorname{curl} A = \overline{B}_j \text{ on } \partial B_{\widetilde{\rho}}(a_j) \right\}$$

for constants $\overline{\psi}_j$ and $\overline{B}_j$. Minimizers $\{u_{min}, A_{min}\}$ have been treated in [2] and [23]. We choose $\widetilde{\rho} = \rho - C\varepsilon^{\alpha_0}$, and the interpolation functions $\{u_{int}, A_{int}\}$ can be chosen as in the proof of Lemma 2.6. A long but straightforward calculation shows both (3.29) and (3.30).

3. By the definition of the renormalized energy and (3.28),

$$\begin{aligned} &\pi W(\{a_j\}) + o_\varepsilon(1) \\ &\quad \le G_\varepsilon(u^\rho_\varepsilon, A^\rho_\varepsilon) - \pi d |\log \varepsilon| \\ &\quad \le G_\varepsilon(u_\varepsilon, A_\varepsilon) - \pi d |\log \varepsilon| + o_\varepsilon(1) \\ &\qquad - \frac{1}{2}\int_{\Omega_\rho} |\nabla |u_\varepsilon||^2 + \left| \frac{j_\varepsilon}{|u_\varepsilon|} - (\nabla \Theta_a + \nabla \psi_\star - A_\star) \right|^2 + |B_\varepsilon - B_\star|^2 \, dx. \end{aligned} \tag{3.31}$$

Our assumption $G_\varepsilon(u_\varepsilon, A_\varepsilon) - \pi d |\log \varepsilon| \le \pi W(\{a_j\}) + \mu$ and (3.31) yield

$$\lim_{\varepsilon \to 0} \int_{\Omega_\rho} |\nabla |u_\varepsilon||^2 \, dx \le 2\mu + o(\varepsilon, \rho), \tag{3.32}$$

$$\lim_{\varepsilon \to 0} \int_{\Omega_\rho} \left| \frac{j_\varepsilon}{|u_\varepsilon|} - (\nabla \Theta_a + \nabla \psi_\star - A_\star) \right|^2 dx \le 2\mu + o(\varepsilon, \rho), \tag{3.33}$$

$$\lim_{\varepsilon \to 0} \int_{\Omega_\rho} |B_\varepsilon - B_\star|^2 \, dx \le 2\mu + o(\varepsilon, \rho). \tag{3.34}$$

So (3.33) and (3.34) imply

$$\limsup_{\varepsilon \to 0} \int_{\Omega_\rho} \left| \frac{j_\varepsilon}{|u_\varepsilon|} - v \right|^2 dx \le 4\mu + 2\int_{\Omega_\rho} |\nabla \psi_\star - \nabla \psi_a + A_\star - A_a|^2 \, dx + o(\varepsilon, \rho),$$
$$\limsup_{\varepsilon \to 0} \int_{\Omega_\rho} |B_\varepsilon - B_a|^2 \, dx \le 4\mu + 2\int_{\Omega_\rho} |B_\star - B_a|^2 \, dx + o(\varepsilon, \rho).$$

Adding these two together yields

$$\begin{aligned} &\limsup_{\varepsilon \to 0} \int_{\Omega_\rho} \left| \frac{j_\varepsilon}{|u_\varepsilon|} - v \right|^2 + |B_\varepsilon - B_a|^2 \, dx \\ &\qquad \le 8\mu + 2\int_{\Omega_\rho} |\nabla \psi_\star - \nabla \psi_a + A_\star - A_a|^2 + |B_\star - B_a|^2 \, dx + o(\varepsilon, \rho). \end{aligned}$$

4. We now show that

$$\int_{\Omega_\rho} |\nabla \psi_\star - \nabla \psi_a + A_\star - A_a|^2 + |B_\star - B_a|^2 \, dx \le \mu + o(\varepsilon, \rho).$$

By the definition of the renormalized energy in the appendix,

$$\frac{1}{2}\int_{\Omega_\rho} |\nabla \Theta_a + \nabla \psi_a - A_a|^2 + |B_a - H_0|^2 \, dx = \pi d \log \frac{1}{\rho} + \pi W(\{a_j\}) - \gamma d + o(\rho). \tag{3.35}$$

Using our initial energy bound, along with an energy bound inside each $B_\rho(a_j)$, we find that our comparison function satisfies

$$\begin{aligned}\int_{\Omega_\rho} g_\varepsilon(u_\varepsilon, A_\varepsilon)dx &= \int_{\Omega_\rho} g_\varepsilon(u_\varepsilon^\rho, A_\varepsilon^\rho)dx \\ &\le \pi W(\{a_j\}) - \gamma d + \mu + o(\rho,\varepsilon) + \pi d \log\frac{1}{\rho},\end{aligned} \tag{3.36}$$

where we used (3.30). Finally,

$$\begin{aligned}\int_{\Omega_\rho} |\nabla\Theta_a + \nabla\psi_\star - A_\star|^2 &= \int_{\Omega_\rho} |\nabla\Theta_a|^2 + |\psi_\star - A_\star|^2\,dx - 2\int_{\partial\Omega} \psi_\star \partial_\nu \Theta_a d\omega \\ &\quad + 2\sum_{j=1}^d \int_{\partial B_\rho(a_j)} \left(\psi_\star - \overline{\psi}_\star^j\right)\partial_\nu\Theta_a + \xi_\star \partial_\tau \Theta_a d\omega.\end{aligned} \tag{3.37}$$

The second line of (3.37) is $o(\rho,\varepsilon)$. Combining (3.35)–(3.37) we get

$$\int_\Omega |\nabla\psi_\star - A_\star|^2 + |B_\star - H_0|^2\,dx \le \int_\Omega |\nabla\psi_a - A_a|^2 + |B_a - H_0|^2\,dx + \mu + o(\varepsilon,\rho).$$

Since ψ_a is harmonic and $\partial_\nu\psi_a = \partial_\nu\psi_\star$ on $\partial\Omega$ along with $\Delta^2\xi_a - \Delta\xi_a + 2\pi\sum_{j=1}^d \delta_{a_j} = 0$ and $\xi_a = \xi_\star = 0$, $\Delta\xi_a = \Delta\xi_\star = H_0$ on $\partial\Omega$, then

$$\begin{aligned}&\int_\Omega |\nabla\psi_\star - \nabla\psi_a + A_\star - A_a|^2 + |B_\star - B_a|^2\,dx \\ &\qquad \le 2\int_\Omega |\nabla\psi_a - A_a|^2 + |B_a - H_0|^2 + \mu + o(\varepsilon,\rho) \\ &\qquad \le \mu + o(\varepsilon,\rho). \qquad \square\end{aligned}$$

4. Vortex motion law for almost-energy-minimizing bounds. We are now in the position to prove the vortex motion law for almost-energy-minimizing bounds. In particular we prove the following.

THEOREM 4.1. *If $\{u_\varepsilon, A_\varepsilon\}$ satisfies the SGL equations* (3.2)–(3.7) *and initial conditions* (3.15)–(3.17)*, then*

$$\frac{d}{dt}a_j(t) = \left(\nabla\psi_j - A_\star\right)(a_j(t)) = -\mathcal{J}\nabla_{a_j} W(\{a_k(t)\}), \tag{4.1}$$

where

$$\mathcal{J} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

and $W(\{a_j\})$ is defined by (A.1)*. $A_\star = -\operatorname{curl}\xi_\star$ satisfies the London equation* (3.23) *and $e^{i\Theta_a + i\psi_\star} = \prod_{j=1}^d \frac{x-a_j}{|x-a_j|}e^{i\psi_\star} = \frac{x-a_j}{|x-a_j|}e^{i\psi_j}$ such that $\Delta\psi_\star = 0$ in Ω and $\partial_\nu\psi_\star = -\partial_\nu\Theta_a$ on $\partial\Omega$.*

We use the momentum equation (3.11) to establish (4.1). Let $m = 1, 2$; then

$$\begin{aligned}\frac{1}{2}\partial_t j_{\varepsilon m} &= -\left(u_{\varepsilon x_m} - iA_{\varepsilon m}u_\varepsilon, u_{\varepsilon x_j} - iA_{\varepsilon j}u_\varepsilon\right)_{x_j} + P_{\varepsilon x_m} \\ &\quad - \frac{|u_\varepsilon|^2}{2}E_\varepsilon + \delta_\varepsilon B_\varepsilon\left(E_\varepsilon \times e_3\right).\end{aligned}$$

Then if $|u_\varepsilon| > 0$,

$$u_{\varepsilon x_m} - iA_{\varepsilon m} = \frac{j_{\varepsilon m}}{|u_\varepsilon|}\frac{iu_\varepsilon}{|u_\varepsilon|} + |u_\varepsilon|_{x_m}\frac{u_\varepsilon}{|u_\varepsilon|}$$

and

$$\begin{aligned}\left(\nabla_{A_{\varepsilon m}} u_\varepsilon, \nabla_{A_{\varepsilon j}} u_\varepsilon\right) &= \frac{\left(j_{\varepsilon m}, j_{\varepsilon j}\right)}{|u_\varepsilon|^2} + |u_\varepsilon|_{x_m}|u_\varepsilon|_{x_j} \\ &= v_m \frac{j_{\varepsilon j}}{|u_\varepsilon|} + v_j \frac{j_{\varepsilon m}}{|u_\varepsilon|} - v_m v_j \\ &\quad + \left(\frac{j_{\varepsilon m}}{|u_\varepsilon|} - v_m\right)\left(\frac{j_{\varepsilon j}}{|u_\varepsilon|} - v_j\right) \\ &\quad + |u_\varepsilon|_{x_m}|u_\varepsilon|_{x_j}.\end{aligned}$$

Since $\|\frac{j_\varepsilon}{|u_\varepsilon|}\|_{L^2_{loc}(\Omega_a)} \leq C$, then there is a weak limit in $L^2[0,T;L^2(\Omega_a)]$, which we denote v. Since $|u_\varepsilon| \to 1$ in $L^2(\Omega_a)$ for a.e. t, then $j_\varepsilon \rightharpoonup v = \nabla\Theta_a + \nabla\psi_\star - A_\star$ and

$$v_j \frac{j_{\varepsilon m}}{|u_\varepsilon|} \rightharpoonup v_j v_m.$$

Therefore,

$$\left(\nabla_{A_\varepsilon} u_\varepsilon \otimes \nabla_{A_\varepsilon} u_\varepsilon\right) \rightharpoonup (v \otimes v) + \nu_j, \tag{4.2}$$

where

$$\left(\frac{j_{\varepsilon j}}{|u_\varepsilon|} - v_j\right)\left(\frac{j_{\varepsilon m}}{|u_\varepsilon|} - v_m\right) + |u_\varepsilon|_j |u_\varepsilon|_m \rightharpoonup \nu_j$$

for a finite, symmetric defect measure, $\nu_j \in \mathcal{M}_+(\Omega_a)$. The failure of strong convergence of $|\frac{j_\varepsilon}{|u_\varepsilon|} - v|_{L^2}$ and $|\nabla|u_\varepsilon||_{L^2}$ accounts for this defect measure. ν is finite on Ω due to Lemma 3.10.

We set $\phi \in C_0^\infty(B_{R_0/2})$ such that $\phi = x$ in B_R. Then the conservation of momentum equation (3.11) yields

$$\begin{aligned}&\int_{B_{R_0/2}} \nabla^\perp\phi \cdot j_\varepsilon\big|_t^{t+k}\, dx \\ &\quad = 2\int_t^{t+k} ds \int_{B_{R_0/2}\setminus B_R} \left(\nabla_{A_\varepsilon} u_\varepsilon \otimes \nabla_{A_\varepsilon} u_\varepsilon\right) : \nabla\nabla^\perp\phi\, dx \\ &\qquad + 2\int_t^{t+k} ds \int_{B_{R_0/2}} \left(\frac{|u_\varepsilon|^2}{2} E_\varepsilon - \delta_\varepsilon B_\varepsilon \left(E_\varepsilon \times e_3\right)\right) \cdot \nabla^\perp\phi\, dx,\end{aligned} \tag{4.3}$$

and using (4.2) we get

$$\begin{aligned}&2\int_t^{t+k} ds \int_{B_{R_0/2}\setminus B_R} \left(\nabla_{A_\varepsilon} u_\varepsilon \otimes \nabla_{A_\varepsilon} u_\varepsilon\right) : \nabla\nabla^\perp\phi\, dx \\ &\qquad \to 2\int_t^{t+k} ds \int_{B_{R_0/2}\setminus B_R} (\mu + v \otimes v) : \nabla\nabla^\perp\phi\, dx,\end{aligned} \tag{4.4}$$

where $\mu \in \mathcal{M}_+(\Omega)$ and $v \otimes v \notin L^1(\Omega)$. We can then examine the second term of (4.4) following [24],

$$\begin{aligned}
&\int_t^{t+k} ds \int_{B_{R_0/2}(a_j(s))\setminus B_R(a_j(s))} (v\otimes v) : \nabla\nabla^\perp \phi \, dx \\
&\quad = \int_t^{t+k} ds \int_{B_{R_0/2}(a_j(s))\setminus B_R(a_j(s))} -v\cdot\nabla v\cdot\nabla^\perp \phi \, dx \\
&\qquad + \int_t^{t+k} ds \int_{\partial B_R(a_j(s))} (v\otimes v) : \left(v\otimes n^\perp\right) \, d\omega \\
&\quad = \int_t^{t+k} ds \int_{\partial B_R(a_j(s))} -\left(v\cdot \nabla v \cdot \nu^\perp\right)(n\cdot x) \, d\omega \\
&\qquad + \int_t^{t+k} ds \int_{\partial B_R(a_j(s))} (v\otimes v) : \left(\nu\otimes n^\perp\right) \, d\omega,
\end{aligned}$$

where $n = (1,0)$ and ν is the normal direction at $\partial B_R(a_j(s))$. Let $(I, II) = (\nabla\psi_j - A_\star)$; then the first integral of the right side becomes

$$\begin{aligned}
&\int_0^{2\pi} \left(a_j^x(t)R + R^2\cos\theta\right)\left[v\cdot\nabla v_1(-\sin\theta) + v\cdot\nabla v_2\cos\theta\right] d\theta \\
&\quad = \int_0^{2\pi} \left(a_j^x(t)R + R^2\cos\theta\right)\left[(I - R^{-1}\sin\theta)(I_x + 2R^{-2}\sin\theta\cos\theta)(-\sin\theta)\right. \\
&\qquad \left. +(II + R^{-1}\cos\theta)(I_y - R^{-2}\cos 2\theta)(-\sin\theta)\right] d\theta \\
&\qquad + \int_0^{2\pi} \left(a_j^x(t)R + R^2\cos\theta\right)\left[(I - R^{-1}\sin\theta)(II_x - R^{-2}\cos 2\theta)\cos\theta\right. \\
&\qquad \left. +(II + R^{-1}\cos\theta)(II_y - 2R^{-2}\sin\theta\cos\theta)(\cos\theta)\right] d\theta \\
&\quad = -I\int_0^{2\pi} 2\left(\sin\theta\cos\theta\right)^2 d\theta - I\int_0^{2\pi}\cos^2\theta\cos 2\theta d\theta + O(R) \\
&\quad = -I\int_0^{2\pi}\cos^2\theta = -\pi I.
\end{aligned}$$

If we let $n = (0,1)$, then the integral yields $-\pi II$. Therefore,

$$\frac{d}{dt}a_j = 2\left(\nabla\psi_j - A_\star(a_j)\right) + f_j(\nu),$$

and by (A.10)

$$\frac{d}{dt}a_j = -\mathcal{J}\nabla_{a_j} W(\{a_k(t)\}) + f_j(\nu), \tag{4.5}$$

where

$$\mathcal{J} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Unfortunately, we have little control over how the defect measure affects the vortex motion. In fact the interaction of vortices with any excess energy can be very nontrivial [18].

To finish the proof of Theorem 4.1, we compare the true vortex motion $a_j(t)$ with a solution of the ODE

$$\frac{d}{dt}b_j(t) = -\mathcal{J}\nabla_{b_j}W(\{b_k(t)\})$$

such that $b_j(0) = a_j(0)$. Set

$$\zeta(t) = \sum_{j=1}^{d} |a_j(t) - b_j(t)|;$$

hence $\zeta(0) = 0$. We wish to show $\zeta(t) \equiv 0$ for all $t \in [0,T]$. Take a small time interval so that $\zeta(t) \le t_\delta$. Then

(4.6)
$$\begin{aligned}
\frac{d}{dt}\zeta(t) &\le \sum_{j=1}^{d}\left|\frac{d}{dt}a_j(t) - \frac{d}{dt}b_j(t)\right| \\
&= \sum_{j=1}^{d}\left|\frac{d}{dt}a_j(t) + \mathcal{J}\nabla_{a_j}W(\{a_k\}) - \mathcal{J}\nabla_{a_j}W(\{a_k\}) + \mathcal{J}\nabla_{b_j}W(\{b_k\})\right| \\
&\le \sum_{j=1}^{d}\left|\frac{d}{dt}a_j(t) + \mathcal{J}\nabla_{a_j}W(\{a_k\})\right| + \sum_{j=1}^{d}\left|\mathcal{J}\nabla_{b_j}W(\{b_k\}) - \mathcal{J}\nabla_{a_j}W(\{a_k\})\right| \\
&\le \sum_{j=1}^{d}\left|\frac{d}{dt}a_j(t) + \mathcal{J}\nabla_{a_j}W(\{a_k\})\right| + C\zeta(t).
\end{aligned}$$

As before, consider the time interval $[t, t+k]$, with k small, and the ball $B_R = B_R(a_j(t))$ inside $B_{R_0/2}$. Then set $\phi \in C_0^\infty(B_{R_0/2})$ such that $\phi = x$ in B_R:

(4.7)
$$\begin{aligned}
&\int_{B_{R_0/2}} \nabla^\perp\phi\cdot j_\varepsilon\big|_t^{t+k}\,dx \\
&\quad= 2\int_t^{t+k} ds\int_{B_{R_0/2}\setminus B_R} (\nabla_{A_\varepsilon}u_\varepsilon\otimes\nabla_{A_\varepsilon}u_\varepsilon) : \nabla\nabla^\perp\phi\,dx \\
&\qquad+ 2\int_t^{t+k} ds\int_{B_{R_0/2}}\left(\frac{|u_\varepsilon|^2}{2}E_\varepsilon - B_\varepsilon\,(E_\varepsilon\times e_3)\right)\cdot\nabla^\perp\phi\,dx \\
&\quad= 2\int_t^{t+k} ds\int_{B_{R_0/2}\setminus B_R}\left[\left(v\otimes\frac{j_\varepsilon}{|u_\varepsilon|} + \frac{j_\varepsilon}{|u_\varepsilon|}\otimes v - v\otimes v\right) : \nabla\nabla^\perp\phi\right]dx \\
&\qquad+ 2\int_t^{t+k} ds\int_{B_{R_0/2}\setminus B_R}\left[\left(\frac{j_\varepsilon}{|u_\varepsilon|} - v\right)\otimes\left(\frac{j_\varepsilon}{|u_\varepsilon|} - v\right) : \nabla\nabla^\perp\phi\right]dx \\
&\qquad+ 2\int_t^{t+k} ds\int_{B_{R_0/2}\setminus B_R}\left[\nabla|u_\varepsilon|\otimes\nabla|u_\varepsilon| : \nabla\nabla^\perp\phi\right]dx \\
&\qquad+ 2\int_t^{t+k} ds\int_{B_{R_0/2}}\left(\frac{|u_\varepsilon|^2}{2}E_\varepsilon - \delta_\varepsilon B_\varepsilon\,(E_\varepsilon\times e_3)\right)\cdot\nabla^\perp\phi\,dx \\
&\quad= (A)+(B)+(C)+(D).
\end{aligned}$$

We showed that as $\varepsilon \to 0$, (A) will yield $-2\pi\mathcal{J}\nabla_{a_j^x(t)}W(\{a_k(t)\})$. We will control terms (B) and (C) with the Γ-convergence result of section 2. We first show that $(D) \le C\zeta(t)+o_\varepsilon(1)+o_{R_0}(1)$. If we start with an almost-energy-minimizing sequence, then

$$\begin{aligned} G_\varepsilon(u_\varepsilon, A_\varepsilon)(0) &\le \pi d|\log\varepsilon| + W(\{a_j(0)\}) + o_\varepsilon(1) \\ &= \pi d|\log\varepsilon| + W(\{b_j(t)\}) + o_\varepsilon(1) \\ &\le \pi d|\log\varepsilon| + W(\{a_j(t)\}) + C\zeta(t) + o_\varepsilon(1) \end{aligned} \tag{4.8}$$

by the Lipschitz continuity of $W(x)$, $|W(a)-W(b)| \le C|a-b| = C\zeta$. Therefore,

$$\begin{aligned} \delta_\varepsilon \int_0^t \int_\Omega E_\varepsilon^2 \, dxdt &\le G_\varepsilon(u_\varepsilon, A_\varepsilon)(0) - G_\varepsilon(u_\varepsilon, A_\varepsilon)(t) \\ &\le W(\{a_j(0)\}) - W(\{a_j(t)\}) + o_\varepsilon(1) \\ &\le W(\{b_j(t)\}) - W(\{a_j(t)\}) + o_\varepsilon(1) \\ &\le C\zeta(t) + o_\varepsilon(1). \end{aligned} \tag{4.9}$$

Noting that $|\nabla\phi| \le R_0^{-1}$, then

$$\begin{aligned} &-\int_t^{t+k}\int_{B_{R_0}} |u_\varepsilon|^2 \, E_\varepsilon \cdot \nabla^\perp\phi \, dxds \\ &\quad= \int_t^{t+k}\int_{B_{R_0}} \left(1-|u_\varepsilon|^2\right) E_\varepsilon \cdot \nabla^\perp\phi \, dxds + \int_t^{t+k}\int_{B_{R_0}} E_\varepsilon \cdot \nabla^\perp\phi \, dxds \\ &\quad= \frac{1}{R_0^2}\int_t^{t+k}\int_{B_{R_0}} \frac{1}{\delta_\varepsilon}\left(1-|u_\varepsilon|^2\right)^2 dxds + \delta_\varepsilon \int_t^{t+k}\int_{B_{R_0}} E_\varepsilon^2 \, dxds \\ &\qquad+ \int_t^{t+k}\int_{B_{R_0}} \phi \operatorname{curl} E_\varepsilon \, dxds \\ &\quad\le \frac{\varepsilon^2|\log\varepsilon|}{\delta_\varepsilon}\frac{C}{R_0^2} + C\zeta(t) + \int_t^{t+k}\int_{B_{R_0}} \partial_t\left(\phi B_\varepsilon\right) dxds, \end{aligned} \tag{4.10}$$

where we control the second term of (4.10) by (4.9). To control the third term, note that $\|B_\varepsilon\|_{W^{1,r}} \le C$ uniformly for all $r \in [1,2)$ and $t \in [0, t_\delta]$. Then

$$\begin{aligned} &\int_t^{t+k}\int_{B_{R_0}} \partial_t\left(\phi B_\varepsilon\right) dxds \\ &\quad= \int_{B_{R_0}} \phi B_\varepsilon\big|_t^{t+k} \, dx \\ &\quad\le C|\phi|_{C_0(\Omega)}\left(\|B_\varepsilon(t+k)\|_{W^{1,r}}^{1/p} + \|B_\varepsilon(t)\|_{W^{1,r}}^{1/p}\right) R_0^{2/q}. \end{aligned} \tag{4.11}$$

Combining (4.10) and (4.11) along with (3.1) yields

$$\int_t^{t+k}\int_{B_{R_0}} |u_\varepsilon|^2 \, E_\varepsilon \cdot \nabla^\perp\phi \, dxds = C\zeta(t) + o_\varepsilon(1) + o_{R_0}(1), \tag{4.12}$$

which controls the first term of (D). It should be noted that we first let $\varepsilon \to 0$ and

then let $R \to 0$. The second term of (D) is much simpler:

$$\begin{aligned}&\int_t^{t+k}\int_{B_{R_0}} \delta_\varepsilon B_\varepsilon \left(E_\varepsilon \times e_3\right)\cdot \nabla^\perp \phi \, dxds\\ &= \delta_\varepsilon \int_t^{t+k}\int_{B_{R_0}} E_\varepsilon^2 \, dxds + \delta_\varepsilon \int_t^{t+k}\int_{B_{R_0}} B_\varepsilon^2 \, dxds\\ &\leq C\zeta(t) + o_\varepsilon(1),\end{aligned} \tag{4.13}$$

which finishes control of (D).

Next, we bound terms (B) and (C). From (4.8) and Lemma 3.10

$$\limsup_{\varepsilon\to 0}\left[G_\varepsilon(u_\varepsilon, A_\varepsilon) - \pi d|\log\varepsilon|\right] \leq \pi W(\{a_j\}) + C\zeta(t) + o(\varepsilon,\rho)$$

for $\zeta(t) > 0$; then for any $r > 0$, there is a constant C independent of ε and R such that

$$\begin{aligned}&\limsup_{\varepsilon\to 0}\left\|\frac{j_\varepsilon}{|u_\varepsilon|} - v\right\|^2_{L^2\left(\Omega\setminus\bigcup_{j=1}^d B_R(a_j)\right)} \leq C\zeta(t),\\ &\limsup_{\varepsilon\to 0}\|B_\varepsilon - B_\star\|^2_{L^2\left(\Omega\setminus\bigcup_{j=1}^d B_R(a_j)\right)} \leq C\zeta(t),\\ &\limsup_{\varepsilon\to 0}\|\nabla|u_\varepsilon|\|^2_{L^2\left(\Omega\setminus\bigcup_{j=1}^d B_R(a_j)\right)} \leq C\zeta(t).\end{aligned}$$

Choosing $t_\delta C \leq \zeta(t) \in (0,1)$, then for all $t \in (0, t_\delta)$

$$\begin{aligned}&\limsup_{\varepsilon\to 0}\left\|\frac{j_\varepsilon}{|u_\varepsilon|} - v\right\|_{L^2(B_{R_0/2}\setminus B_R)} \leq C_1\zeta(t),\\ &\limsup_{\varepsilon\to 0}\|B_\varepsilon - B_\star\|_{L^2(B_{R_0/2}\setminus B_R)} \leq C_1\zeta(t),\\ &\limsup_{\varepsilon\to 0}\|\nabla|u_\varepsilon|\|_{L^2(B_{R_0/2}\setminus B_R)} \leq C_1\zeta(t).\end{aligned}$$

This controls $(B)+(C)$. Then sending $\varepsilon \to 0$ yields, for $a = (a^x, a^y)$,

$$LHS \to 2\pi \frac{d}{dt} a_j^x(t)$$

and

$$(I) \to -2\pi \mathcal{J}\nabla_{a_j^x} W(\{a_k(t)\}).$$

Therefore, we find

$$\left|\frac{d}{dt}a_j^x(t) + \mathcal{J}\nabla_{a_j^x} W(\{a_k\})\right| \leq C\zeta(t),$$

and performing a similar estimate for $a_j^y(t)$, we get

$$\left|\frac{d}{dt}a_j(t) + \mathcal{J}\nabla_{a_j} W(\{a_k\})\right| \leq C\zeta(t)$$

and the inequality

$$\frac{d}{dt}\zeta(t) \le C\zeta(t)$$

with $\zeta(0) = 0$, which implies $\zeta \equiv 0$ and the vortex motion law. This finishes Theorem 4.1.

Remark 4.2. Although the SGL equations are dissipative for fixed ε, in the $\varepsilon \to 0$ limit, energy is conserved. This asymptotic behavior strongly depends on the choice of δ_ε in (3.3). For a different rate of $\delta_\varepsilon \to 0$, the system will not conserve energy, as dissipation will dominate.

4.1. Incompressible Euler equations. In [24] the authors were able to show the convergence of the supercurrent to a set of incompressible Euler equations for the Gross–Pitaevskii equation (1.5). To do so they found that the defect measure ν, arising from the limit of the convective term

$$(\nabla u_\varepsilon \otimes \nabla u_\varepsilon) \rightharpoonup (v \otimes v + \nu),$$

is curl-free, and that allowed $\operatorname{div} \nu$ to be written as the gradient of a distribution, and hence pushed into the pressure term. It is reasonable to ask whether the supercurrent equation (3.11) converges weakly to a set of Euler equations. Although $j_\varepsilon \rightharpoonup v$ is divergence-free, there are a number of difficulties controlling (3.11), including the lack of a curl-free supercurrent ($\operatorname{curl} v = -B_\star \neq 0$) and the loss of control over the term $\frac{1}{2}|u_\varepsilon|^2 E_\varepsilon$ in $\Omega \backslash \bigcup_{j=1}^d B_r(a_j)$. In the end it may be possible to study only the vorticity equation

$$\frac{1}{2}\operatorname{curl}\left(\partial_t j_\varepsilon + |u_\varepsilon|^2 E_\varepsilon\right) = -\operatorname{curl}\operatorname{div}\left(\nabla_{A_\varepsilon} u_\varepsilon \otimes \nabla_{A_\varepsilon} u_\varepsilon\right) - \left(\delta_\varepsilon \operatorname{div} E_\varepsilon\right) B_\varepsilon,$$

which has better control on both sides of the equation.

Appendix. Renormalized energy. For completeness we include a discussion of the renormalized energy for the full GL energy functional, which we need to verify the dynamic law SGL equations. We note the analysis is similar to [1, 2, 23]. We aim to prove two theorems that characterize both the renormalized energy and the gradient of the renormalized energy.

A.1. Renormalized energy.

THEOREM A.1. *The renormalized energy* $W\left(\{a_j\}_1^d\right)$ *obeys the system*

$$\begin{aligned} W(a_1, \dots, a_d) = &-\pi \sum_{i \neq j} \log|a_i - a_j| - \pi \sum_{j=1}^d R(a_j) + \pi \sum_{j=1}^d \xi(a_j) \\ &+ \frac{H_0^2}{2}|\Omega| - \frac{1}{2} H_0 \int_{\partial\Omega} \partial_\nu \xi \, d\omega. \end{aligned} \tag{A.1}$$

Here ξ *satisfies* $A = -\operatorname{curl}\xi$ *and*

$$\begin{aligned} -\Delta^2 \xi + \Delta \xi &= 2\pi \sum_{j=1}^d \delta_{a_j} \text{ in } \Omega, \\ \xi &= 0 \text{ on } \partial\Omega, \\ \Delta \xi &= H_0 \text{ on } \partial\Omega, \end{aligned} \tag{A.2}$$

and

$$R(x) = P(x) - \sum_{j=1}^{d} \log|x - a_j|, \tag{A.3}$$

where

$$\begin{aligned} \Delta P &= 2\pi \sum_{j=1}^{d} \delta_{a_j} \ \textit{in } \Omega, \\ P &= 0 \ \textit{on } \partial\Omega. \end{aligned} \tag{A.4}$$

We will first prove Theorem A.1 by using a combination of arguments found in [1, 2, 23]. We first note that $u_\star \in \mathbb{S}^1$ can be written as $u_\star = e^{i\Theta_a + i\psi_\star} = \prod_{j=1}^{d} \frac{x-a_j}{|x-a_j|} e^{i\psi_\star}$ for harmonic function $\psi_\star$. $\psi_\star$ is difficult to study, as it is a multivalued function. Since $\partial_\nu u = 0$ on $\partial\Omega$, we can find the conjugate harmonic function P such that

$$u_\star \times \nabla u_\star = \begin{pmatrix} -\partial_2 P \\ \partial_1 P \end{pmatrix} = -\operatorname{curl} P.$$

Then the boundary condition implies $0 = \nu \cdot \nabla^\perp P = \partial_\tau P$ or P is constant on $\partial\Omega$. We choose $P = 0$ to simplify the discussion below. Therefore, the equation for the congugate harmonic equation becomes

$$\begin{aligned} \Delta P_a &= 2\pi\delta_a \text{ in } \Omega, \\ P_a &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{A.5}$$

In two dimensions the singularity at a is $O(\log)$ so we can define

$$P_a(x) = \log|x - a| + S_a(x), \tag{A.6}$$

which is no longer a multivalued function (unlike Θ_a). Furthermore, $S_a(x)$ is harmonic and defined everywhere in Ω. We now outline the proof of the theorem by first defining the class on which we define the renormalized energy. To calculate $W(\{a_j\})$ we subtract the self-induction energy from a canonical harmonic map, and what is left is a function of the d vortex locations. We set $\Omega_\rho = \Omega \setminus \bigcup_{j=1}^{d} B_\rho(a_j)$ and

$$\mathcal{H}^1(a_j, \rho) = \begin{cases} u \in H^1\left(\Omega_\rho, \mathbb{S}^1\right), \ A \in H^1\left(\Omega, \mathbb{R}^2\right) \text{ such that} \\ u = \dfrac{x - a_j}{|x - a_j|} \text{ on } \partial B_\rho(a_j) \quad \text{and} \quad \partial_\nu u = 0 \text{ on } \partial\Omega, \\ \operatorname{div} A = 0 \text{ in } \Omega \quad \text{and} \quad \nu \cdot A = 0 \quad \text{and} \quad \operatorname{curl} A = H_0 \text{ on } \partial\Omega. \end{cases}$$

We set

$$E_\delta(u, A) = \frac{1}{2} \int_{\Omega_\delta} |\nabla_A u|^2 \, dx + \frac{1}{2} \int_{\Omega} |\operatorname{curl} A - H_0|^2 \, dx \tag{A.7}$$

and

$$\mu_\delta = \inf_{(u,A) \in \mathcal{H}^1(a_j, \rho)} E_\delta(u, A). \tag{A.8}$$

Claim A.2. *We have that μ_δ is achieved, and for $\delta < \delta_0$ we have $\mu_\delta \le \pi d \log \frac{1}{\delta} + C$, where $C = C(\{a_j\}, \delta_0)$.*

Proof. This is proved by creating suitable comparison functions $(v_\varepsilon, B_\varepsilon)$, following [2], such that $G_\varepsilon(v_\varepsilon, B_\varepsilon) \le \pi d|\log\varepsilon| + C$. We fix d distinct points $a_1, \dots, a_d$ and R small enough such that $B_R(a_j) \subset \Omega$ and $B_R(a_i) \cap B_R(a_j) = \emptyset$. To construct our comparison functions we start with

$$w_0 = v_\varepsilon = e^{i\theta_a} = \frac{x - a_j}{|x - a_j|} \quad \text{on } \partial B_R(a_j),$$
$$B_0 = B_\varepsilon = (iw_0, \nabla w_0)$$

outside the $B_R(a_j)$'s. We define a mollifier ζ such that $\zeta = 1$ for $r \ge 1$ and $\zeta = 0$ for $\zeta \le 1/2$. We can now define the comparison functions inside the $B_R(a_j)$'s.

$$v_\varepsilon = \frac{x - a_j}{|x - a_j|}\zeta\left(\frac{x - a_j}{\varepsilon}\right),$$
$$B_\varepsilon = B_\varepsilon\left(\frac{x - a_j}{|x - a_j|}R\right)\zeta\left(\frac{x - a_j}{R}\right).$$

A simple computation shows that $\int_{B_R} g_\varepsilon(v_\varepsilon, B_\varepsilon)dx \le C$ and

$$\int_{B_R(a_j)} g_\varepsilon(v_\varepsilon, B_\varepsilon)dx \le \pi|\log\varepsilon| + C.$$

Then we can use the direct method of calculus of variations to establish the existence of μ_δ. ∎

Claim A.3. *We have for $\delta < \delta_0$ and for a minimizer (v_δ, B_δ) of (A.7)*

$$\int_\Omega |\mathrm{curl} B_\delta - H_0|^2\, dx \le C$$

and

$$\int_{\Omega_\delta} |\nabla_{B_\delta} v_\delta|^2\, dx \ge \pi d \log\frac{1}{\delta} - C$$

for $C = C(\{a_j\}, \delta_0)$.

Proof. We again look to [2] for guidance. Let (v_δ, B_δ) be a minimizer of (A.7) such that $\operatorname{div} B_\delta = 0$ with $\nu \cdot B_\delta = 0$ on $\partial\Omega$. Then there exists a ξ_δ such that $B_\delta = \operatorname{curl}\xi_\delta$, where $\Delta\xi_\delta = h_\delta = \operatorname{curl} B_\delta$, where $\xi_\delta = 0$ on $\partial\Omega$. Then

$$\int_{\Omega_\delta} |\nabla_{B_\delta} v_\delta|^2\, dx = \int_{\Omega_\delta} |\nabla v_\delta|^2 + |\nabla\xi_\delta|^2 + 2\,[\xi_\delta, v_\delta]\, dx,$$

where we have $[\xi_\delta, v_\delta] = (iv_\delta, \nabla v_\delta) \times \nabla\xi_\delta$. Then

$$\int_{\Omega_\delta} (iv_\delta, \nabla v_\delta) \times \nabla\xi_\delta\, dx = \int_{\Omega_\delta} \xi_\delta\, \mathrm{curl}(iv_\delta, \nabla v_\delta)dx + \sum_{j=1}^{d} \int_{\partial B_\delta(b_j)} \xi_\delta(iv_\delta, \partial_\tau v_\delta)d\omega$$
$$= \sum_{j=1}^{d} \int_{\partial B_\delta(b_j)} \xi_\delta(iv_\delta, \partial_\tau v_\delta)d\omega,$$

where τ is tangential vector to $\partial B_\delta(b_j)$ since $v_\delta \in \mathbb{S}^1$. We show that

$$\int_{\Omega_\delta} |\nabla_{B_\delta} v_\delta|^2\, dx \ge \pi d \log\frac{1}{\delta} - C.$$

Then the top follows. ∎

CLAIM A.4. *Let ξ be a solution to*

$$-\Delta^2\xi + \Delta\xi = 2\pi\sum_{j=1}^{d}\delta_{a_j} \text{ in } \Omega,$$

$$\xi = 0 \text{ on } \partial\Omega,$$

$$\Delta\xi = H_0 \text{ on } \partial\Omega.$$

Then $\xi_\delta \to \xi$ in $W^{2,2}(\Omega)$ as $\delta \to 0$, where $B_\delta = -\operatorname{curl}\xi_\delta = -\nabla^\perp\xi_\delta$.

Proof. By the minimality in the class,

$$-\Delta^2\xi_\delta + \Delta\xi_\delta = 0 \text{ in } \Omega,$$
$$\Delta\xi = 0 \text{ on } \partial\Omega.$$

Following an argument similar to [2] we establish the above. ∎

CLAIM A.5. *Let*

$$\overline{\mu}_\delta = \min\left\{\frac{1}{2}\int_{\Omega_\delta}|\nabla u|^2 dx, u \in H^1(\Omega_\delta, \mathbb{S}^1), \deg(u, \partial B_\delta(a_i)) = 1, \partial_\nu u = 0 \text{ on } \partial\Omega\right\}.$$

Then

$$\left|\frac{1}{2}\int_{\Omega_\delta}|\nabla v_\delta|^2\,dx - \overline{\mu}_\delta\right|^2 \to 0 \tag{A.9}$$

as $\delta \to 0$.

Proof. By standard elliptic estimates we find

$$\frac{1}{2}\int_{\Omega_\delta}|\nabla v_\delta|^2\,dx \geq \overline{\mu}_\delta + o(\delta),$$

and using $R(\delta) \to 0$ we decouple the phase terms from the magnetic field such that

$$E_\delta(v_\delta, B_\delta) = \frac{1}{2}\int_{\Omega_\delta}|\nabla v_\delta|^2 dx + \frac{1}{2}\int_\Omega |\nabla\xi|^2 + |\Delta\xi - H_0|^2 dx + 2\pi\sum_{j=1}^{d}\xi(a_j) + o(1);$$

see [2]. ∎

We can now use the analysis in [1] to characterize the form of the phase terms.

CLAIM A.6. *Let $1 < p < 2$; then the map v_δ remains bounded in $W^{1,p}$ and $v_\delta \to v$ in $W^{1,p}$ to $v = e^{i\Theta_a + ik} = \prod_{j=1}^{d}\frac{x-a_j}{|x-a_j|}e^{i\psi_\star}$, where $\psi_\star$ is harmonic and*

$$\partial_\nu\psi_\star = -\left(\prod_{j=1}^{d}\frac{x-a_j}{|x-a_j|}, \partial_\nu\prod_{j=1}^{d}\frac{x-a_j}{|x-a_j|}\right) \quad \text{on } \partial\Omega.$$

Proof. Note that v_δ takes values in $\mathbb{S}^1$, and therefore

$$\operatorname{curl}(iv_\delta, \nabla v_\delta) = 0 \in \Omega_\delta.$$

Let P be a solution to

$$\Delta P = 2\pi \sum_{j=1}^{d} \delta_{a_j} \text{ in } \Omega_\delta,$$
$$P = 0 \text{ in } \partial\Omega_\delta.$$

Then

$$\operatorname{curl}((iv_\delta, \nabla v_\delta) + \nabla P) = 0.$$

Therefore, there exists H_δ such that

$$(iv_\delta, \nabla v_\delta) + \nabla P = \operatorname{curl} H_\delta.$$

Following the analysis in [1] we get the claim. □

Proof of Theorem A.1. We show, for any configuration $\{a_j\} = \{a_1, \dots, a_d\}$,

$$\mu_\delta(a_j) = W(\{a_j\}) + \pi d \log \frac{1}{\delta} + o(1),$$

where $W(\{a_j\})$ is defined by (A.1)–(A.4).

From (A.9)

$$\mu_\delta(b_j) = \frac{1}{2}\int_{\Omega_\delta} |\nabla v_\delta|^2 dx + \frac{1}{2}\int_\Omega |\nabla\xi|^2 + |\Delta\xi - H_0|^2 dx + 2\pi \sum_{j=1}^{d} \xi(a_j) + o(1).$$

But (A.2) gives

$$\frac{1}{2}\int_\Omega |\nabla\xi|^2 + |\Delta\xi|^2 dx = -\pi \sum_{j=1}^{d} \xi(a_j) + \frac{H_0}{2}\int_{\partial\Omega} \partial_\nu \xi \, d\omega.$$

Then

$$\mu_\delta(a_j) = \frac{1}{2}\int_{\Omega_\delta} |\nabla v_\delta|^2 dx + \pi \sum_{j=1}^{d} \xi(a_j) - \frac{H_0}{2}\int_{\partial\Omega} \partial_\nu \xi d\omega + \frac{H_0^2}{2}|\Omega| + O(\delta) + o(1),$$

but from [1, 2]

$$\frac{1}{2}\int_{\Omega_\delta} |\nabla v_\delta|^2 dx = \pi d \log \frac{1}{\delta} + \omega(\{a_j\}, d, H_0) + O(\delta).$$

Then

$$W(\{a_j\}) = \omega(\{a_j\}, d, H_0) + \pi \sum_{j=1}^{d} \xi(a_j) - \frac{H_0}{2}\int_{\partial\Omega} \partial_\nu \xi \, d\omega + \frac{H_0^2}{2}|\Omega|.$$

We now use the canonical form of the phase, (A.6).

$$\omega(\{a_j\}, d, H_0) = -\pi \sum_{j=1}^{d} \log|a_i - a_j| - \pi \sum_{j=1}^{d} R(a_j) + \frac{1}{2}\int_{\partial\Omega} P \partial_\nu P \, d\omega,$$

where $R(x) = P(x) - \sum_{j=1}^{d} \log|x - a_j|$, and by previous discussion we know that $P = 0$ on $\partial\Omega$. Therefore, we get (A.1). □

A.2. Renormalized energy gradient. We now establish the form of the gradient of the renormalized energy. In this subsection we derive two forms of the gradient. Theorem A.7 is used in section 4.

THEOREM A.7. *Let* $u_\star = e^{i\Theta_a + i\psi_\star} = \sum_{j=1}^d \frac{x-a_j}{|x-a_j|} e^{i\psi_\star}$ *satisfy* $\Delta\psi_\star = 0$ *in* Ω *and* $\partial_\nu \psi_\star = -\partial_\nu \Theta_a$*; then* $e^{i\Theta_a + i\psi_\star} = \frac{x-a_j}{|x-a_j|} e^{i\psi_j}$ *defines* ψ_j*. Let* ξ *satisfy* $\Delta^2\xi - \Delta\xi + 2\pi \sum_{j=1}^d \delta_{a_j}$ *in* Ω *and* $\xi = 0$, $\Delta\xi = H_0$ *on* $\partial\Omega$*. If* $W(\{a_j\})$ *is as defined by* (A.1)–(A.4)*, then*

$$
\begin{aligned}
DW(\{a_j\}) &= 2\pi \left[\left(-\frac{\partial \psi_1}{\partial x_2}(a_1) + \frac{\partial \xi}{\partial x_1}(a_1), \frac{\partial \psi_1}{\partial x_1}(a_1) + \frac{\partial \xi}{\partial x_2}(a_1) \right), \dots, \right. \\
&\qquad \left. \left(-\frac{\partial \psi_d}{\partial x_2}(a_d) + \frac{\partial \xi}{\partial x_1}(a_d), \frac{\partial \psi_d}{\partial x_1}(a_d) + \frac{\partial \xi}{\partial x_2}(a_d) \right) \right] \\
&= 2\pi \left[\left(-\frac{\partial \psi_1}{\partial x_2}(a_1) + A_2(a_1), \frac{\partial \psi_1}{\partial x_1}(a_1) - A_1(a_1) \right), \dots, \right. \\
&\qquad \left. \left(-\frac{\partial \psi_d}{\partial x_2}(a_d) + A_2(a_d), \frac{\partial \psi_d}{\partial x_1}(a_d) - A_1(a_d) \right) \right].
\end{aligned} \tag{A.10}
$$

Proof. We will use [1] for inspiration. Fix all vortices except vortex a_j, which we call y. Therefore, we have

$$
\begin{aligned}
\Delta P &= 2\pi \sum_{i \neq j} \delta_{a_i} + 2\pi \delta_y \text{ in } \Omega, \\
P &= 0 \text{ on } \partial\Omega,
\end{aligned}
$$

and

$$
\begin{aligned}
-\Delta^2 \xi + \Delta \xi &= 2\pi \sum_{i \neq j} \delta_{a_i} + 2\pi \delta_y \text{ in } \Omega, \\
\Delta \xi &= H_0 \text{ on } \partial\Omega, \\
\xi &= 0 \text{ on } \partial\Omega.
\end{aligned} \tag{A.11}
$$

Set

$$
\begin{aligned}
\Psi(x,y) &= P(x,y) - \sum_{i \neq j} \log|x - a_i|, \\
R(x,y) &= \Psi(x,y) - \log|x-y|.
\end{aligned}
$$

Then the following equations hold:

$$
\begin{aligned}
\Delta \Psi &= 2\pi \delta_y \text{ in } \Omega, \\
\Psi &= -\sum_{i \neq j} \log|x - a_i| = h(x) \text{ on } \partial\Omega.
\end{aligned}
$$

Then for $a, \tilde{a} \in \Omega$

$$
\begin{aligned}
2\pi \left(\Psi(a,\tilde{a}) - \Psi(\tilde{a},a) \right) &= \int_{\partial\Omega} \Psi(\sigma,\tilde{a}) \partial_\nu \Psi(\sigma,a) - \Psi(\sigma,a) \partial_\nu \Psi(\sigma,\tilde{a}) d\sigma \\
&= \int_{\partial\Omega} h(\sigma) \left(\partial_\nu \Psi(\sigma,a) - \partial_\nu \Psi(\sigma,\tilde{a}) \right) d\sigma.
\end{aligned}
$$

Therefore, by the symmetry of the $\log|x-y|$

$$2\pi\left(R(a,\tilde{a})-R(\tilde{a},a)\right)=\int_{\partial\Omega}h(\sigma)\left(\partial_\nu P(\sigma,a)-\partial_\nu P(\sigma,\tilde{a})\right)d\sigma.$$

Now set

$$\zeta(x)=P(x,\tilde{a})-P(x,a)=R(x,\tilde{a})-R(x,a)+\log\frac{|x-\tilde{a}|}{|x-a|}.$$

Then

$$\begin{aligned}\Delta\zeta&=2\pi\left(\delta_{\tilde{a}}-\delta_a\right)\text{ in }\Omega,\\ \zeta&=0\text{ on }\partial\Omega.\end{aligned}$$

Multiplying by $\sum_{i\neq j}\log|x-a_i|$ and integrating yields

$$2\pi\sum_{i\neq j}\zeta(a_i)+\int_{\partial\Omega}\partial_\nu\zeta(\sigma)\sum_{i\neq j}\log|\sigma-a_i|\,d\sigma=2\pi\sum_{i\neq j}\log\frac{|\tilde{a}-a_i|}{|a-a_i|},$$

which gives

$$\begin{aligned}2\pi\left(R(a,\tilde{a})-R(\tilde{a},a)\right)&=-\int_{\partial\Omega}\left(\sum_{i\neq j}\log|\sigma-a_i|\right)\partial_\nu\zeta(\sigma)d\sigma\\&=2\pi\sum_{i\neq j}\zeta(a_i)+2\pi\sum_{i\neq j}\log\frac{|\tilde{a}-a_i|}{|a-a_i|}.\end{aligned}$$

Thus for $\tilde{a}$ fixed and varying a we get $2\pi\left(R_x(a,\tilde{a})-R_y(\tilde{a},a)\right)=2\pi\sum_{i\neq j}R_y(a_i,a)$ or

$$2\pi\left(R_x(a,a)-R_y(a,a)\right)=2\pi\sum_{i\neq j}R_y(a_i,a). \tag{A.12}$$

We now concern ourselves with (A.11). If we let $y=\tilde{a}$, then multiplying by $\xi(\sigma,a)$ and integrating over Ω yields

$$\begin{aligned}&2\pi\xi(a,\tilde{a})+2\pi\sum_{i\neq j}\xi(a_i,a)\\&\quad=H_0\int_{\partial\Omega}\partial_\nu\xi(\sigma,a)d\sigma-\int_\Omega\Delta\xi(\sigma,a)\Delta\xi(\sigma,\tilde{a})+\nabla\xi(\sigma,a)\nabla\xi(\sigma,\tilde{a})d\sigma,\end{aligned}$$

which implies

$$\begin{aligned}&2\pi\left(\xi(a,\tilde{a})-\xi(\tilde{a},a)\right)+2\pi\sum_{i\neq j}\left(\xi(a_i,a)-\xi(a_i,\tilde{a})\right)\\&\qquad=H_0\int_{\partial\Omega}\left(\partial_\nu\xi(\sigma,a)-\partial_\nu\xi(\sigma,\tilde{a})\right)d\sigma.\end{aligned}$$

Then for $\tilde{a}$ fixed and varying a we get

$$2\pi\left(\xi_x(a,a)-\xi_y(a,a)\right)+2\pi\sum_{i\neq j}\xi_y(a_i,a)=H_0\int_{\partial\Omega}\partial_\nu\xi_y(\sigma,a)d\sigma. \tag{A.13}$$

Therefore, noting that we can write the renormalized energy as

$$W(a) = -\pi \sum_{i\neq j} \log|a-a_i| - \pi \sum_{k\neq l\neq j} \log|a_k - a_l| - \pi \sum_{i\neq j} R(a_i,a) - \pi R(a,a) + \pi \sum_{i\neq j} \xi(a_i,a) + \pi\xi(a,a) + \frac{H_0^2}{2}|\Omega| - \frac{H_0}{2}\int_{\partial\Omega} \partial_\nu \xi(\sigma,a)d\sigma,$$

then

$$W_a(a) = -\pi \sum_{i\neq j} \frac{a-a_i}{|a-a_i|^2} - \pi \sum_{i\neq j} R_y(a_i,a) - \pi R_x(a,a) - \pi R_y(a,a) + \pi \sum_{i\neq j} \xi_y(a_i,a) + \xi_x(a,a) + \xi_y(a,a) - \frac{H_0}{2}\int_{\partial\Omega} \partial_\nu \xi_y(\sigma,a)d\sigma. \tag{A.14}$$

Using (A.12) and (A.13) in (A.14) gives

$$W_a(a) = -\pi \sum_{i\neq j} \frac{a-a_i}{|a-a_i|^2} - 2\pi R_x(a,a) + 2\pi\xi_x(a,a). \tag{A.15}$$

So if

$$\begin{aligned} -S_j(x) + \xi(x,a) &= -P(x,a) + \log|x-a| + \xi(x,a) \\ &= -R(x,a) - \pi \sum_{i\neq j} \log|x-a_i| + \xi(x,a), \end{aligned}$$

and since $\nabla\psi_j = -\nabla^\perp S_j$, then (A.15) becomes

$$W_a(a) = -2\pi\nabla S_j(a) + 2\pi\nabla\xi(a_j) = \begin{pmatrix} -\partial_2\psi_j(a) + A_2(a) \\ \partial_1\psi_j(a) - A_1(a) \end{pmatrix}.$$

Since ψ_j is the harmonic conjugate to $R(x,a)$, we establish Theorem A.7. ∎

A.3. Solution in the disk $B_R(0)$. We start with a configuration of d vortices at $a_j(0)$. Let us redefine ξ as our $\xi \longmapsto \xi + \chi$ such that

$$-\Delta^2\xi + \Delta\xi = 2\pi \sum_{j=1}^{d} \delta_{a_j(t)}$$

in $B_R(0)$, $\Delta\xi = \xi = 0$ on $\partial B_R(0)$,

$$-\Delta^2\chi + \Delta\chi = 0$$

in $B_R(0)$, $\Delta\chi = H_0$ and $\chi = 0$ on $\partial B_R(0)$,

$$\Delta P = 2\pi \sum_{j=1}^{d} \delta_{a_j(t)}$$

in $B_R(0)$, and $P = 0$ on $\partial B_R(0)$. For the solid disk, $B_R(0)$, we can solve ξ, χ, and P. Let $a_j = \rho_j e^{i\psi_j}$; then, if we use a general method for calculating Green's functions

[27], the associated Green's functions are

$$\begin{aligned}
\xi(r,\phi) &= -\sum_{j=1}^{d} \log \sqrt{\frac{R^2 - 2r\rho_j\cos(\phi-\psi_j) + \left(\frac{r\rho_j}{R}\right)^2}{r^2 - 2r\rho_j\cos(\phi-\psi_j)+\rho_j^2}} \\
&\quad + \sum_{j=1}^{d} K_0\left(\sqrt{r^2 - 2r\rho_j\cos(\phi-\psi_j)+\rho_j^2}\right) - \frac{I_0(\rho_j)I_0(r)K_0(R)}{I_0(R)} \\
&\quad - 2\sum_{j=1}^{d}\sum_{n=1}^{\infty} \frac{I_n(\rho_j)I_n(r)K_n(R)}{I_n(R)}\cos(n(\phi-\psi_j)), \\
\chi(r,\phi) &= H_0\left(\frac{I_0(r)}{I_0(R)} - 1\right), \\
\Phi(r,\phi) &= \sum_{j=1}^{d} \log \sqrt{\frac{r^2 - 2r\rho_j\cos(\phi-\psi_j)+\rho_j^2}{R^2 - 2r\rho_j\cos(\phi-\psi_j) + \left(\frac{r\rho_j}{R}\right)^2}},
\end{aligned}$$

where $I_n(r)$ and $K_n(r)$ are the modified Bessel functions of the first and second kinds.

Acknowledgments. This research was completed as part of a dissertation at the Courant Institute. The author would like to thank his advisor F.-H. Lin, who suggested this problem and provided valuable guidance. The author would also like to thank R.V. Kohn, J. Shatah, S. Gustafson, F. Han, and R. Jerrard for many interesting and helpful discussions. Finally, the author appreciates the comments of an anonymous referee that greatly improved the final draft.

REFERENCES

[1] F. Bethuel, H. Brezis, and F. Helein, *Ginzburg-Landau Vortices*, Birkhäuser Boston, Boston, MA, 1994.
[2] F. Bethuel and T. Riviere, *Vortices for a variational problem related to superconductivity*, Ann. Inst. H. Póincare Anal. Non Linéaire, 12 (1995), pp. 243–303.
[3] H. Brezis and T. Gallouët, *Nonlinear Schrödinger evolution equations*, Nonlinear Anal., 4 (1980), pp. 677–681.
[4] C. Caroli and K. Maki, *Fluctuations of the order parameter in Type*-II *superconductors*. I. *Dirty Limit*, Phys. Rev., 159 (1967), pp. 306–315.
[5] C. Caroli and K. Maki, *Fluctuations of the order parameter in Type*-II *superconductors*. II. *Pure Limit*, Phys. Rev., 159 (1967), pp. 315–326.
[6] Y. M. Chen and F. H. Lin, *Evolution of harmonic maps with Dirichlet boundary conditions*, Comm. Anal. Geom., 1 (1993), pp. 327–346.
[7] A. Chorin, *Vorticity and Turbulence*, Springer-Verlag, New York, 1994.
[8] J. E. Colliander and R. L. Jerrard, *Ginzburg-Landau vortices: Weak stability and Schrödinger equation dynamics*, J. Anal. Math., 77 (1999), pp. 129–205.
[9] M. Comte and P. Mironescu, *The behavior of a Ginzburg-Landau minimizer near its zeroes*, Calc. Var. Partial Differential Equations, 4 (1996), pp. 323–340.
[10] B. S. Deaver and W. M. Fairbank, *Experimental evidence for quantized flux in superconducting cylinders*, Phys. Rev. Lett., 7 (1961), pp. 51–53.
[11] R. Doll and M. Näbauer, *Experimental proof of magnetic flux quantization in a superconducting ring*, Phys. Rev. Lett., 7 (1961), pp. 43–46.
[12] Q. Du *Global existence and uniqueness of solutions of the time-dependent Ginzburg-Landau model for superconductivity*, Appl. Anal., 53 (1994), pp. 1–17.
[13] W. E, *Dynamics of vortices in Ginzburg-Landau theories with applications to superconductivity*, Phys. D, 77 (1994), pp. 383–404.
[14] R. Feynman, *Statistical Mechanics* W.A. Benjamin, Reading, MA, 1972.
[15] V. L. Ginzburg and L. D. Landau, *On the theory of superconductivity*, in Collected Papers of L. D. Landau, Pergamon Press, New York, 1965, pp. 626–633.

[16] S. Gustafson and D. Spirn, *Vortex Motion Law for the Maxwell-Higgs Equations*, preprint.
[17] R. L. Jerrard, *Lower bounds for generalized Ginzburg–Landau functionals*, SIAM J. Math. Anal., 30 (1999), pp. 721–746.
[18] R. L. Jerrard, *personal communication.*
[19] R. L. Jerrard and H. M. Soner, *Dynamics of Ginzburg-Landau vortices*, Arch. Ration. Mech. Anal., 142 (1998), pp. 99–125.
[20] F. H. Lin, *Some dynamic properties of Ginzburg-Landau vortices*, Comm. Pure Appl. Math., 49 (1996), pp. 323–359.
[21] F. H. Lin, *Complex Ginzburg-Landau equations and dynamics of vortices, filaments, and codimension-*2 *submanifolds*, Comm. Pure Appl. Math., 51 (1998), pp. 385–441.
[22] F. H. Lin, *Vortex dynamics for the nonlinear wave equation*, Comm. Pure Appl. Math., 52 (1999), pp. 737–761.
[23] F. H. Lin and Q. Du, *Ginzburg-Landau vortices: Dynamics, pinning, and hysteresis*, SIAM J. Math. Anal., 28 (1997), pp. 1265–1293.
[24] F. H. Lin and J. X. Xin, *On the incompressible fluid limit and the vortex motion law of the nonlinear Schrödinger equation*, Comm. Math. Phys., 200 (1999), pp. 249–274.
[25] M. Machida and H. Kaburaki, *Direct simulation of the time-dependent Ginzburg-Landau equation for type-*II *superconducting thin film: Vortex dynamics and* V-I *characteristics*, Phys. Rev. Lett., 71 (1993), pp. 3206–3209.
[26] E. Madelung, *Quantentheorie in hydrodynamischer form*, Z. Physik, 40 (1927), p. 322.
[27] Y. A. Melnikov, *Green's Functions in Applied Mathematics*, Computational Mechanics Publications, Southampton, 1995.
[28] J. Neu, *Vortices in complex scalar fields*, Phys. D, 43 (1990), pp. 385–406.
[29] L. Peres and J. Rubinstein, *Vortex dynamics in* $U(1)$ *Ginzburg-Landau models*, Phys. D, 64 (1993), pp. 299–309.
[30] E. Sandier and S. Serfaty, *Global minimizers for the Ginzburg-Landau functional below the first critical magnetic field*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 119–145.
[31] S. Serfaty, *Local minimizers for the Ginzburg-Landau equations near critical magnetic field: Part* I, Comm. Cont. Math., 2 (1999), pp. 213–254.
[32] S. Serfaty, *Local minimizers for the Ginzburg-Landau equations near critical magnetic field: Part* II, Comm. Cont. Math., 3 (1999), pp. 295–333.
[33] D. Spirn, *Vortex dynamics of the full time-dependent Ginzburg-Landau equations*, Comm. Pure Appl. Math., 55 (2002), pp. 537–581.
[34] M. Struwe, *On the asymptotic behavior of minimizers of the Ginzburg-Landau model in* 2 *dimensions*, Differential Integral Equations, 7 (1994), pp. 1613–1624.
[35] M. Tinkham, *Introduction to Superconductivity* McGraw-Hill, New York, 1996.
[36] A. Tonomura, H. Kasai, O. Kamimura, T. Matsuda, K. Harada, J. Shimoyama, K. Kishio, and K. Kitazawa, *Motion of vortices in superconductors*, Nature, 397 (1999), pp. 308–309.

SIAM J. MATH. ANAL.
Vol. 34, No. 6, pp. 1477–1502

DIFFRACTIVE NONLINEAR GEOMETRIC OPTICS FOR SHORT PULSES*

DEBORAH ALTERMAN† AND JEFFREY RAUCH‡

Abstract. This paper considers the behavior of pulse-like solutions of length $\varepsilon \ll 1$ to semilinear systems of hyperbolic partial differential equations on the time scale $t = O(1/\varepsilon)$ of diffractive geometric optics. The amplitude is chosen so that nonlinear effects influence the leading term in the asymptotics.

For pulses of larger amplitude so that the nonlinear effects are pertinent for times $t = O(1)$, accurate asymptotic solutions lead to transport equations similar to those valid in the case of wave trains (see [D. Alterman and J. Rauch, *J. Differential Equations*, 178 (2002), pp. 437–465]). The opposite is true here. The profile equation for pulses for $t = O(1/\varepsilon)$ is different from the corresponding equation for wave trains.

Formal asymptotics leads to equations for a leading term in the expansion and for correctors. The equations for the correctors are in general not solvable, being plagued by small divisor problems in the continuous spectrum. This makes the construction of accurate approximations subtle. We use low-frequency cutoffs depending on ε to avoid the small divisors.

Key words. pulses, diffraction, geometric optics, short wavelength asymptotics, hyperbolic partial differential equations

AMS subject classifications. 35B40, 35C20, 35L05, 35Q60

PII. S0036141002403584

1. Introduction.

The simplest pulse-like solutions arise as plane wave solutions of constant coefficient homogeneous hyperbolic equations.

1.1. Linear plane waves.

Assumption 1.1 (symmetric hyperbolicity).

$$L(\partial_y) = \partial_t + \sum_{j=1}^{d} A_j \partial_{x_j}, \tag{1.1}$$

where the coefficients A_j are constant $N \times N$ hermitian symmetric matrices.

The space-time variable is

$$y = (t, x) \in \mathbb{R}^{1+d} \qquad \text{with dual variables} \qquad (\tau, \xi).$$

If $f : \mathbb{R} \to \mathbb{C}^N$ is smooth and $\beta = (\tau, \xi)$, then the chain rule yields

$$L\, f(y.\beta) = L(\beta) f'(y.\beta).$$

Thus $L(f(y.\beta)) = 0$ when f' takes values in the nullspace of $L(\beta)$.

*Received by the editors March 5, 2002; accepted for publication (in revised form) November 18, 2002; published electronically May 29, 2003.

http://www.siam.org/journals/sima/34-6/40358.html

†Department of Applied Mathematics, University of Colorado, Boulder, CO 80309 (dalterman@earthlink.net). The research of this author was partially supported by the U.S. National Science Foundation under grant NSF-DMS-9810751.

‡Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (rauch@umich.edu). The research of this author was partially supported by the U.S. National Science Foundation under grants NSF-DMS-9500823 and NSF-INT-9314095.

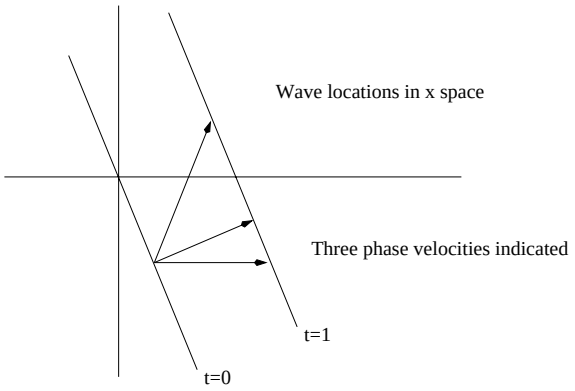


FIG. 1.1. *Planar fronts and phase velocities.*

Recall that the *characteristic variety*, $Char\, L$, is the set of $(\tau, \xi) \in \mathbb{R}^{1+d} \setminus 0$ satisfying the *dispersion relation* $\det L(\tau, \xi) = 0$. The characteristic variety is a real conic algebraic variety. For $\beta \in Char\, L$ one has the orthogonal decomposition

$$\mathbb{C}^N \;=\; \ker L(\beta) \;\oplus_\perp\; \text{range}\, L(\beta).$$

DEFINITION 1.2. *For $\beta \in Char\, L$, $\pi = \pi(\beta)$ is the orthogonal projection of $\mathbb{C}^N$ onto $\ker L(\beta)$. Define the partial inverse $Q(\beta)$ by*

$$Q\pi = 0, \qquad QL(\beta) = (I - \pi).$$

Then $u = f(y.\beta)$ is a *plane wave solution* of $L\,u = 0$ when $\beta \in Char\, L$ and f satisfies the polarization $\pi(\beta)\, f = f$.

1.2. Plane pulses and group velocity. If, in addition,

$$f(s) \to 0 \qquad \text{as} \qquad s \to \pm\infty,$$

then the family of plane wave solutions

$$u^\varepsilon \;:=\; f\Big(\frac{y.\beta}{\varepsilon}\Big)$$

describes pulses with planar wave fronts. If f has compact support, then the pulse u^ε is supported in an $O(\varepsilon)$ neighborhood of the hyperplane $y.\beta = 0$. The pulse cross section is given by the function $f^\varepsilon(s) := f(s/\varepsilon)$. The function f is called the *profile* of this pulse family. For profiles which tend to zero as $s \to \pm\infty$, the conditions $\pi(\beta) f = f$ and $\pi(\beta) f' = f'$ are equivalent.

At $t = 0$ (resp., $t = 1$) the pulse is supported near the planes $x.\xi = 0$ (resp., $x.\xi = -\tau$). This is indicated in Figure 1.1.

The phase is given by

$$y.\beta = t\tau + x.\xi = (x - \mathbf{v}t).\xi$$

for any velocity vector $\mathbf{v}$ satisfying

$$\mathbf{v}.\xi = -\tau.$$

For any such $\mathbf{v}$, the pulse family is given by

$$h^\varepsilon(x - \mathbf{v}t), \qquad \text{where} \qquad h^\varepsilon(x) := f^\varepsilon(x.\xi) = f(x.\xi/\varepsilon).$$

The pulse family can be viewed as moving with any one of these *phase velocities*. Three such velocities are sketched in Figure 1.1. In dimension $d > 1$, the phase velocity is not uniquely determined.

In contrast Definition 1.4 below shows that the *group velocity* is well defined at smooth points β of the characteristic variety.

For $\xi \neq 0$ the points $(\tau, \xi) \in Char\, L$ which project to ξ are those points such that $-\tau$ is a real eigenvalue of $\sum \xi_j A_j$. Thanks to the hyperbolicity assumption this is a finite and nonempty set of points for each ξ, and so the variety has codimension 1. As such its singular points form a variety of codimension at least 2 so that most points of the variety are smooth in the sense of the next assumption.

Assumption 1.3 (smooth point of the characteristic variety). $\beta = (\tau_0, \xi_0)$ belongs to the characteristic variety, and there is a conic neighborhood of ξ_0 and a real analytic function $\tau(\xi)$ on that neighborhood so that on a conic neighborhood of β the variety is given by the equation $\tau = \tau(\xi)$.

DEFINITION 1.4. *For β a smooth point of the characteristic variety, the* group velocity *is defined by*

$$\textit{group velocity} \;:=\; \mathbf{v} \;:=\; -\nabla_\xi \tau(\xi_0).$$

Since τ is homogeneous of degree 1 in ξ, the Euler homogeneity relation implies that

$$\xi.\nabla_\xi \tau(\xi) = \tau(\xi).$$

This implies that the group velocity satisfies $\mathbf{v}.\xi = -\tau$, the equation defining phase velocities. *The group velocity is the correct choice from among the possible phase velocities.*

1.3. Wave trains versus pulses. The geometric optics approximations which are most familiar concern the short wavelength limit of wave trains (see [24]). Wave trains and pulses are contrasted in Figure 1.2. Standard geometric optics yields equations for the envelope of wave trains. The methods go under the name of the slowly varying envelope approximation (SVEA) in science journals. A rule of thumb is that to use the SVEA the amplitude should not change more than 10% per wavelength. The wave train in Figure 1.2 is a borderline case for this rule. The rule of thumb suggests that one must have about ten to twenty wavelengths per pulse length before the SVEA is a reliable approximation.

For much shorter pulses like the one on the right in Figure 1.2 the SVEA is clearly inappropriate. Interest in short-pulse phenomena, which violate this slowly varying envelope assumption, has increased with the development of ultrafast lasers which produce few-cycle pulses. Rothenberg [25] clearly described the problems arising from treating short pulses as wave trains. Short-pulse solutions have been studied via full numerical simulation as in [28], [15], [16], [17], and [14]. A variety of asymptotic attacks are proposed and pursued in [13], [10], [21], [20], and [22]. In this paper the equation defining the leading order asymptotics is simple, and the approximation is proved to be accurate in the limit of small wavelength. Thus, from the above list only those which are consistent with our approximation can also be accurate. Only those whose equations are as simple can be competitive. It is our evaluation that with

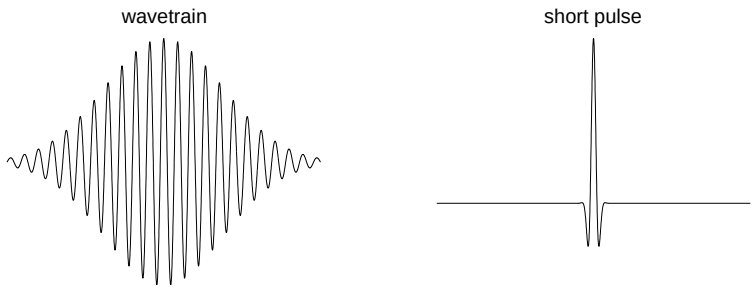


FIG. 1.2. *Example of a wave train and a short pulse.*

these two criteria in mind, a strong case can be made for our approach, but out of politeness we leave detailed comparison to the interested reader.

A difficulty in the study of pulses and wave trains is that the terms in the hyperbolic equation are of radically different magnitudes. Simply dropping the smaller ones and keeping the large ones usually leads to completely incorrect results. This is apparent in the simplest families of short pulses in $\mathbb{R}^{1+1}$,

$$u^\varepsilon := e^{-t} f((x-t)/\varepsilon), \qquad \partial_t u^\varepsilon + \partial_x u^\varepsilon + u^\varepsilon = 0.$$

The three terms of the equation are of sizes $O(1/\varepsilon)$, $O(1/\varepsilon)$, and $O(1)$. Dropping the relatively small $O(1)$ term yields the approximate solution $f((x-t)/\varepsilon)$, which is completely inaccurate.

A second difficulty in short-pulse asymptotics is that formally imitating the expansions of geometric optics generates equations for a leading term and correctors in an asymptotic expansion. Generically, the equations for the correctors cannot be solved. For the time scale $t = O(1)$, before the onset of diffractive effects, Yoshikawa [26], [27] showed that if one imposes physically unnatural assumptions guaranteeing that the corrector equations can be solved, then one does get an accurate description. In [5] we showed that the leading term is accurate without the unnatural assumption and extended the construction to the case of curved wavefronts. In a sequence of articles Carles and Rauch [7], [8], [9] studied the passage of spherical pulse solutions of semilinear wave equations across focal points.

For pulses on the scale of diffractive geometric optics, [1] includes some of the results of the present paper but notably does not prove a rate of convergence of the error as $\varepsilon \to 0$. There is also a study of pulses from Lannes' perspective of waves of broad spectrum (spectre large) in Barrailh and Lannes [6]. It is likely that the present analysis can be extended to nearly planar wavefronts as in the work of Dumas [12] for the diffractive wave train case. The present article contains the proofs of results described and used in [2] and [3].

Typical analytic expressions for the wave forms in Figure 1.2 are

$$\text{wave train :} \qquad a(x)\, e^{ix_1/\varepsilon} \qquad \text{with Fourier transform} \qquad \hat{a}(\xi - (1/\varepsilon), 0);$$

$$\text{pulse :} \qquad a(x_1/\varepsilon) b(x') \qquad \text{with Fourier transform} \qquad \varepsilon \hat{a}(\varepsilon\xi_1)\, \hat{b}(\xi').$$

The Fourier transform of the wave train is localized near $(1/\varepsilon, 0)$, which is called the carrier frequency in applications.

The Fourier transform of the pulse is spread over a box of dimensions $1/\varepsilon \times 1$ in (ξ_1, ξ') space. There is no carrier frequency. There is no exponential prefactor which

renders the quotient slowly varying. Some of the asymptotic approaches cited above are flawed because they insist on identifying a carrier frequency.

The approximations take the form

$$\begin{array}{lllll} \text{wave train :} & U(y\,,\, y.\beta/\varepsilon) & \text{with} & U(y\,,\,\theta) \ \text{ periodic in } \theta; \\ \text{pulse :} & U(y\,,\, y.\beta/\varepsilon) & \text{with} & U(y\,,\, z) \to 0 \ \text{ as } z \to \infty. \end{array}$$

In both cases $\beta \in Char\, L$.

In the latter case the function $U(y, \cdot)$ represents the profile of the pulse. In the former case it gives the envelope of the wave train. The pulse approximation can be called the *slowly varying profile approximation* since the profiles vary on the scale $O(1)$, which is much longer than the pulse length $O(\varepsilon)$.

1.4. The basic problem. Consider the behavior for $t \sim 1/\varepsilon$ of solutions to a system of equations

$$L(\partial_y)u^\varepsilon + \Phi(u^\varepsilon) = 0, \qquad u^\varepsilon(0,x) = \varepsilon^p f\left(x, \frac{x.\xi_0}{\varepsilon}\right), \tag{1.2}$$

where $\beta = (\tau_0, \xi_0)$ is a smooth point of the characteristic variety.

Assumption 1.5 (short pulse initial data). The function $f(x,z)$ satisfies

$$\forall\, N, \quad \langle \xi, \zeta\rangle^N \hat{f}(\xi,\zeta) \ \in \ L^\infty(\mathbb{R}^{d+1}). \tag{1.3}$$

This assumption is slightly stronger than the assumption $f(x,z) \in H^s(\mathbb{R}^{d+1}_{x,z})$ for all $s > 0$, used in [1] and [2], but is weaker than the Schwartz class. We will see that if one starts with f in the Schwartz class, then generically the pulse profile will not be Schwartz class for $t > 0$.

Assumption 1.6 (order J nonlinearity). The nonlinear function $\Phi(u)$ is of order $J \geq 2$ in the sense that for all $|\alpha| \leq J-1$, $\partial^\alpha \Phi(0) = 0$. Denote by $\Phi_J(u)$ the homogeneous Taylor polynomial of degree J approximating $\Phi(u)$ near $u = 0$.

Assumption 1.7 (magnitude of the solution). The exponent p is chosen so that $p = 1/(J-1)$. This insures that nonlinear effects become important on the time scale $t = O(1/\varepsilon)$.

To see that for waves of this amplitude it is reasonable that the nonlinear term is pertinent for times $t = O(1/\varepsilon)$ and not before, make the following back-of-an-envelope estimate. The nonlinear term is of size $\varepsilon^{pJ} = \varepsilon^{p+1}$. The accumulated effect of the nonlinear term for times $t = O(1/\varepsilon)$ is crudely estimated as

$$\frac{1}{\varepsilon}\, \varepsilon^{p+1} \ = \ \varepsilon^p.$$

Since ε^p is the size of our solution it is reasonable to expect the accumulated nonlinear effects to be important on these time scales.

Assumption 1.8 (polarization). The initial data f satisfy the polarization condition $\pi(\beta) f(x,z) = f(x,z)$.

Definition 1.9. *Define the scalar real second order homogeneous differential operator $P(\partial_x)$ by*

$$P(\partial_x) := -\frac{1}{2} \sum_{l,m=1}^{d} \left.\frac{\partial^2 \tau}{\partial \xi_l \partial \xi_m}\right|_{\xi=\xi_0} \frac{\partial^2}{\partial x_l \partial x_m}. \tag{1.4}$$

With these assumptions and definitions, the approximate pulse-like solutions have the form

$$u^{\varepsilon}_{\text{approx}} = \varepsilon^p U_0\Big(\varepsilon t, t, x, \frac{\tau_0 t + \xi_0.x}{\varepsilon}\Big), \qquad \lim_{|z|\to\infty} U_0(T,t,x,z) = 0. \tag{1.5}$$

The slowly varying profile U_0 is polarized as usual, $\pi(\beta)U_0 = U_0$, and is determined from its initial data by the pair of evolution equations

$$(\partial_t + \mathbf{v}.\partial_x)\, U_0 = 0, \qquad \partial_{Tz} U_0 + P(\partial_x)U_0 + \pi(\beta)\partial_z \Phi_J(U_0) = 0.$$

The second equation, for which $T = 0$ is characteristic, is the pulse version of the nonlinear Schrödinger equation.

As these are the key equations that need to be solved in order to understand the behavior of solutions to (1.2), we pause briefly to discuss them. The first equation is handled by writing

$$U_0(T,t,x,z) \;=\; \mathrm{U}_0(T, x - \mathbf{v}t, z). \tag{1.6}$$

The second equation is then equivalent to

$$\partial_{Tz}\mathrm{U}_0 + P(\partial_x)\mathrm{U}_0 + \pi(\beta)\partial_z\Phi_J(\mathrm{U}_0) = 0. \tag{1.7}$$

On the face of it, this is a differential equation in the $d+2$ variables T, x, z. In Proposition 4.1, it is shown that τ'' has rank $\le d-1$, so the differential operator has derivatives in at most $d+1$ independent directions.

To see that (1.7) gives rise to a well-defined evolution, write it formally as

$$\partial_T \mathrm{U}_0 + \partial_z^{-1} P(\partial_x)\mathrm{U}_0 + \pi(\beta)\Phi_J(\mathrm{U}_0) = 0.$$

The operator $\partial_z^{-1}P(\partial_x)$ is antisymmetric on the H^s, which for s large are invariant under Φ_J. Corollary 4.12 implies that for f as above, there is a $T_* \in\,]0,\infty]$ and a unique

$$\mathrm{U}_0 \;\in\; C\big([0,T_*[\,; \cap_s H^s(\mathbb{R}^{d+1}_{x,z})\big)$$

satisfying (1.7) and the initial condition $\mathrm{U}_0|_{T=0} = f$. If $T_* < \infty$, then for all $s > (d+1)/2$,

$$\lim_{T\to T_*} \big\|\mathrm{U}_0(T)\big\|_{H^s(\mathbb{R}^{d+1}_{x,z})} \;=\; \infty.$$

Having constructed U_0, define an approximate solution by (1.6) and (1.5). Our main theorem asserts that the error in this approximation tends to zero as $\varepsilon \to 0$. To motivate a class of natural norms to measure this error, note that $u^{\varepsilon}_{\text{approx}} = O(\varepsilon^p)$. Differentiating u^{ε} costs a power of $1/\varepsilon$ but no worse, so one has

$$\big(\varepsilon\partial\big)^{\alpha} u^{\varepsilon}_{\text{approx}} \;=\; O(\varepsilon^p). \tag{1.8}$$

Denote by $\mathbb{V}$ the $(d+1)$-dimensional space of constant coefficient vector fields. Choose a basis $V_1, \dots, V_d$ of the d-dimensional subspace of fields which are tangent to the hyperplane $\{y.\beta = 0\}$. Choose the basis so that $V_1, \dots, V_{d-1}$ are tangent to $\{t = 0\}$. Then these $d-1$ vectors are a basis for the constant fields on $\mathbb{R}^d_x$ which are

tangent to $\{x.\xi_0 = 0\}$. Differentiating in the d directions $V_1, \dots, V_d$ does not bring out a factor of $1/\varepsilon$, and one has

$$\big(V_1, \dots, V_d\big)^\alpha u^\varepsilon_{\text{approx}} \;=\; O(\varepsilon^p). \tag{1.9}$$

Choose a $(d+1)$st field W, which completes the $V_1, \dots, V_d$ to a basis of $\mathbb{V}$. Since $\xi \neq 0$, this vector field can be chosen tangent to $\{t = 0\}$. Define

$$V_{d+1} \;= \arctan(y.\beta)\, W. \tag{1.10}$$

This vector field vanishes on $\{y.\beta = 0\}$ and so is tangent to that hyperplane. Any smooth vector field tangent to this hyperplane is a linear combination of the V_j with smooth coefficients. Any smooth vector field on $\mathbb{R}^d_x$ tangent to $\{x.\xi_0 = 0\}$ is a combination of $V_1, \dots, V_{d-1}, V_{d+1}|_{\{t=0\}}$ with smooth coefficients. One also has $V_{d+1} u^\varepsilon_{\text{approx}} = O(\varepsilon^p)$. Summarizing, one has for all $\alpha \in \mathbb{N}^{2(d+1)}$

$$\big(\varepsilon\partial_y, V_1, \dots, V_d, V_{d+1}\big)^\alpha u^\varepsilon_{\text{approx}} = O(\varepsilon^p)\,. \tag{1.11}$$

The next result is a straightforward consequence of our main result, Theorem 8.1. Note the technical point that the derivation V_d is not permitted in the error estimate.

THEOREM 1.10. *With the notation of the previous paragraphs, for any $\underline{T} < T_*$ there is an $\varepsilon_0 > 0$ so that for $0 < \varepsilon < \varepsilon_0$ problem (1.2) has a smooth solution $u^\varepsilon \in C^\infty\big(\{0 \le T \le \underline{T}/\varepsilon\} \times \mathbb{R}^d\big)$. The solution is well approximated by $u^\varepsilon_{\text{approx}}$ in the sense that for all $\alpha \in \mathbb{N}^{2(d+1)}$ there is a $C = C(\alpha)$ so that*

$$\big\|\big(\varepsilon\partial_y, V_1, \dots, V_{d-1}, V_{d+1}\big)^\alpha \big(u^\varepsilon - u^\varepsilon_{\text{approx}}\big)\big\|_{L^\infty([0,\underline{T}/\varepsilon]\times\mathbb{R}^d)} \le C\varepsilon^{p+\min(1/5,p)}. \tag{1.12}$$

Remark. Using the techniques of [5], one can show that there is a different family of exact solutions $u^\varepsilon_{\text{ex}}$ with error estimate including V_d, that is,

$$\big\|\big(\varepsilon\partial_y, V_1, \dots, V_{d-1}, V_d, V_{d+1}\big)^\alpha \big(u^\varepsilon_{\text{ex}} - u^\varepsilon_{\text{approx}}\big)\big\|_{L^\infty([0,\underline{T}/\varepsilon]\times\mathbb{R}^d)} \le C\varepsilon^{p+\min(1/5,p)}. \tag{1.13}$$

The initial data of the new family are small perturbations of the initial data for the family u^ε.

2. Formal asymptotics. Seek approximate solutions to the initial value problems with short-pulse initial data,

$$L\,u^\varepsilon + \Phi(u^\varepsilon) = 0, \qquad u^\varepsilon(0,x) = \varepsilon^p\, f\left(x, \frac{x.\xi_0}{\varepsilon}\right), \qquad p = \frac{1}{J-1}. \tag{2.1}$$

The initial function f is assumed to satisfy Assumptions 1.5 and 1.8 for a β satisfying Assumption 1.3. J is the order of the nonlinearity as in Assumption 1.6.

Motivated by its success in the analogous situation of wave train solutions for which f is periodic in z, a first attempt is to try to find a profile

$$U(\varepsilon, T, y, z) \;\sim\; \sum_{j=0}^{\infty} \varepsilon^j\, U_j(T, y, z) \;=\; \sum_{j=0}^{\infty} \varepsilon^j\, U_j(T, t, x, z), \tag{2.2}$$

where $U_j \to 0$ as $|z| \to \infty$, and

$$u^\varepsilon \;\sim\; \varepsilon^p U\left(\varepsilon, \varepsilon t, y, \frac{y.\beta}{\varepsilon}\right). \tag{2.3}$$

The chain rule implies that a sufficient condition guaranteeing that u^ε defined by $u^\varepsilon = U(\varepsilon, \varepsilon t, y, y.\beta/\varepsilon)$ satisfies the differential equation $Lu^\varepsilon + \Phi(u^\varepsilon) = 0$ is that U satisfy

$$L\Big((\varepsilon\partial_T, 0) + \partial_y + \frac{\beta}{\varepsilon}\,\partial_z\Big)\,\varepsilon^p U(\varepsilon, T, y, z) + \Phi(\varepsilon^p U(\varepsilon, T, y, z)) = 0. \tag{2.4}$$

We pursue the less ambitious strategy, which is to satisfy

$$L\Big((\varepsilon\partial_T, 0) + \partial_y + \frac{\beta}{\varepsilon}\partial_z\Big)\,\varepsilon^p U(\varepsilon, T, y, z) + \Phi(\varepsilon^p U(\varepsilon, T, y, z)) \sim 0 \quad \text{as} \quad \varepsilon \to 0, \tag{2.5}$$

in which case

$$Lu^\varepsilon + \Phi(u^\varepsilon) \;\sim\; 0 \qquad \text{as} \quad \varepsilon \to 0. \tag{2.6}$$

We take U to be a sum of only three terms, in which case the equivalence in (2.5) can be no smaller than $O(\varepsilon^{2p+1})$. Two crucial facts affect our implementation of this strategy:

1. A trio of equations, derived in section 3, determine U_0 from its initial data at $t = T = 0$, namely,

$$\begin{aligned} &\pi(\beta)\, U_0 = U_0, \\ &\partial_t U_0 + \mathbf{v}.\partial_x U_0 = 0, \\ &\partial_T \partial_z U_0 + P(\partial_x) U_0 + \pi(\beta)\,\partial_z \Phi_J(U_0) = 0. \end{aligned} \tag{2.7}$$

The middle transport equation of (2.7) is solved by defining $\mathrm{U}_0(T, x, z)$ as in (1.6) so that

$$\mathrm{U}_0(T, x - \mathbf{v}t, z) = U_0(T, t, x, z). \tag{2.8}$$

2. The equations that one finds for the correctors $U_1, U_2, \ldots$ *are not in general solvable*. These equations involve the operator ∂_z^{-1}, which does not act well on a function whose Fourier transform with respect to z does not vanish at the origin, or equivalently, whose integral with respect to z is nonzero. For most choices of initial data, including those for which the transform vanishes on a neighborhood of zero at time $t = 0$, $\int U_0\, dz$ does not vanish at later times, and hence the equations for the correctors are not solvable.

The second fact is the main difficulty this paper overcomes. In our study [5] of geometric optics before the onset of diffractive effects, a similar problem was encountered. In that case we were able to construct correctors which had a different form than the leading term. In the present case of diffractive geometric optics, we do not know how to find such modified correctors.

A crucial ingredient in the analysis is the representation of the exact solution u^ε in terms of an "exact profile" $\mathcal{V}(\varepsilon, t, x, \phi)$ as in [18] and [19] by setting

$$u^\varepsilon = \varepsilon^p\, \mathcal{V}\Big(\varepsilon, t, x, \frac{x.\xi_0}{\varepsilon}\Big). \tag{2.9}$$

A key difference between $\mathcal{V}$ and U is that in the phase variable slot $x.\xi_0/\varepsilon$ replaces $\beta.y/\varepsilon$ for U. To maintain this distinction, the profile variable for which one inserts $x.\xi_0/\varepsilon$ is called ϕ and the profile variable associated with $\beta.y/\varepsilon$ is z. The chain rule shows that u^ε defined by (2.9) satisfies $Lu^\varepsilon + \Phi(u^\varepsilon) = 0$ when

$$L\Big(\partial_t, \partial_x + \frac{\xi_0}{\varepsilon}\partial_\phi\Big)\,\varepsilon^p \mathcal{V} + \Phi(\varepsilon^p \mathcal{V}) = 0. \tag{2.10}$$

The exact profile $\mathcal{V}(\varepsilon, t, x, \phi)$ is then determined from (2.10) and the initial condition

$$\mathcal{V}(\varepsilon, 0, x, \phi) = f(x, \phi). \tag{2.11}$$

The existence and uniqueness of solutions to the initial value problem formed by (2.10) and (2.11) are examined in section 4. Our error estimates proceed by proving that

$$U_0\Big(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big) - \mathcal{V}(\varepsilon, t, x, \phi) \to 0 \qquad \text{as } \varepsilon \to 0. \tag{2.12}$$

To establish (2.12), first note that the chain rule implies that if (2.5) holds, then as $\varepsilon \to 0$,

$$L\Big(\partial_t, \partial_x + \frac{\xi_0}{\varepsilon}\partial_\phi\Big)\varepsilon^p U\Big(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big) + \Phi\Big(\varepsilon^p U\Big(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big)\Big) \sim 0. \tag{2.13}$$

Thus, if one has correctors $U_1, U_2, \ldots$ to the leading profile, then $U(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi)$ defines an accurate approximate solution of the equation for $\mathcal{V}$, and (2.12) would follow.

The difficulty is the absence of correctors—equivalently, the fact that we do not get (2.5). To circumvent this problem, we solve nearby problems with low-frequency cutoffs applied to the nonlinear term and the initial data. The cutoff problems propagate the property of having a Fourier transform with respect to z, which vanishes on a neighborhood of the origin.

Choose a cutoff function $\chi(\zeta) \in C^\infty(\mathbb{R})$ which vanishes for $|\zeta| < 1$ and is identically equal to 1 for $|\zeta| \geq 3/2$. Define

$$\chi^\delta := \chi^\delta(D_z) = \mathcal{F}_z^{-1}\chi(\zeta/\delta)\mathcal{F}_z,$$

where $\mathcal{F}_z$ denotes the Fourier transform in z. Seek

$$U^\delta(\varepsilon, T, y, z) = U_0^\delta(T, y, z) + \varepsilon U_1^\delta(T, y, z) + \varepsilon^2 U_2^\delta(T, y, z) \tag{2.14}$$

as an approximate solution of the cutoff equation

$$\begin{aligned} L\Big((\varepsilon\partial_T, 0) + \partial_y + \frac{\beta}{\varepsilon}\partial_z\Big)\varepsilon^p U^\delta(\varepsilon, T, y, z) + \chi^\delta(D_z)\,\Phi(\varepsilon^p U^\delta(\varepsilon, T, y, z)) \\ = O(\varepsilon^{2p+1}) \end{aligned} \tag{2.15}$$

with initial data

$$U^\delta(0, 0, x, z) = \chi^\delta(D_z) f(x, z). \tag{2.16}$$

Then the main result is proved by showing that

$$U_0^\delta(T, t, x, z) - U_0(T, t, x, z) = O(\delta), \tag{2.17}$$

and that for $\delta = \varepsilon^{0.4}$,

$$U^\delta\Big(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big) - \mathcal{V}(\varepsilon, t, x, \phi) \;=\; O\big(\varepsilon^{\min\{p\,,\,1/5\}}\big). \tag{2.18}$$

The proof has three main steps. First the approximate solution U^δ satisfying (2.15) and (2.16) is constructed in Proposition 4.3 and Corollary 4.12. Proposition 5.1 proves the convergence of U_0^δ to U_0 as in (2.17). Then Proposition 7.1 proves the error estimate (2.18). Propositions 5.1 and 7.1 combine to yield (2.12).

3. Derivation of the profile equations for U_j^δ. In this section we analyze (2.15). The construction of the correctors U_1^δ and U_2^δ works for $\delta > 0$. For $\delta = 0$ the analysis shows that the construction of such an approximate solution is in general not possible.

Define U^δ by (2.14). By convention set $U_{-1}^\delta = U_{-2}^\delta = U_3^\delta = U_4^\delta = 0$. Then computing the left-hand side of (2.15) yields

$$\varepsilon^p\Big(\frac{1}{\varepsilon}L(\beta)\partial_z + L(\partial_y) + \varepsilon\partial_T\Big)(U_0^\delta + \varepsilon U_1^\delta + \varepsilon^2 U_2^\delta) + \chi^\delta(D_z)\Phi(\varepsilon^p U^\delta).$$

Grouping by powers of ε yields

$$\sum_{j=-1}^{j=3} \varepsilon^{p+j}\ \left\{\partial_T U_{j-1}^\delta + L(\partial_y)\,U_j^\delta + L(\beta)\,\partial_z U_{j+1}^\delta\right\} + \varepsilon^{p+1}\chi^\delta\Phi_J(U_0^\delta)$$
$$+\chi^\delta\left\{\Phi(\varepsilon^p U^\delta) - \Phi_J(\varepsilon^p U_0^\delta)\right\}. \tag{3.1}$$

We use formula (2.14) for times $t = O(1/\varepsilon)$. Thus if $U_1^\delta(T,t,x,z)$ grew linearly as $t \to \infty$, the term εU_1 would become as large as the U_0 term and would no longer be a small corrector. In order to represent a small correction it is necessary that U_1^δ grow sublinearly as $t \to \infty$. In fact, the correctors will be uniformly bounded in t, x. Sublinearity will play a crucial role in the derivation of the equations satisfied by the U_j.

3.1. Annihilating the ε^{p-1} term. This term is equal to

$$\varepsilon^{p-1}\,L(\beta)\partial_z U_0^\delta. \tag{3.2}$$

It is annihilated by imposing the polarization

$$U_0^\delta = \pi(\beta)U_0^\delta \tag{3.3}$$

from Definition 1.2. This is consistent with the polarization imposed on the initial data $f(x,z)$ in Assumption 1.8.

3.2. Annihilating the ε^p term. This term is equal to

$$\varepsilon^p\ \left\{L(\partial_y)\,U_0^\delta + L(\beta)\,\partial_z U_1^\delta\right\}. \tag{3.4}$$

To annihilate (3.4), one annihilates in turn its image under $\pi(\beta)$ and then its image under $Q(\beta)$.

Multiplying (3.4) by $\pi(\beta)$ eliminates the U_1^δ term since $\pi(\beta)L(\beta) = 0$. Using the polarization of U_0^δ from (3.3), one finds

$$\pi(\beta)L(\partial_y)\pi(\beta)\,U_0^\delta = 0.$$

As shown in [11], whenever β is a smooth point of the characteristic variety,

$$\pi(\beta)L(\partial_y)\pi(\beta) = \big(\partial_t + \mathbf{v}.\partial_x\big)\pi(\beta).$$

This identity yields the transport equation

$$\partial_t U_0^\delta + \mathbf{v}.\partial_x U_0^\delta = 0. \tag{3.5}$$

Setting $Q(\beta)$ times the ε^p term equal to zero yields

$$(I-\pi(\beta))\,\partial_z U_1^\delta = -Q(\beta)L(\partial_y)U_0^\delta. \tag{3.6}$$

This is the key troublesome equation for the correctors. In order for the equation to be solvable in $H^s(\mathbb{R}^{d+1}_{x,z})$, one needs

$$\frac{1}{\zeta}\,\mathcal{F}_z\Big(Q(\beta)\,L(\partial_y)U_0^\delta\Big)$$

to be square integrable near $\zeta = 0$. There is no reason to expect that this condition will be satisfied when $\delta = 0$.

However, (3.11) below defining U_0^δ for $\delta > 0$ implies that the Fourier transform of U_0^δ with respect to z vanishes on a neighborhood of $\zeta = 0$ as soon as it does so at $T = 0$. Thus we can solve (3.6) to find

$$(I-\pi(\beta))\,U_1^\delta = -(\partial_z)^{-1}Q(\beta)L(\partial_y)U_0^\delta \qquad \text{when} \quad \delta > 0. \tag{3.7}$$

3.3. Annihilating the ε^{p+1} term. This term is equal to

$$\varepsilon^{p+1}\Big(\partial_T U_0^\delta + L(\partial_y)\,U_1^\delta + L(\beta)\,\partial_z U_2^\delta + \chi^\delta \Phi_J(U_0^\delta)\Big). \tag{3.8}$$

When $\delta > 0$, use (3.7) to write

$$U_1^\delta = \pi(\beta)U_1^\delta + (I-\pi(\beta))U_1^\delta = \pi(\beta)U_1^\delta - (\partial_z)^{-1}Q(\beta)L(\partial_y)U_0^\delta.$$

Then setting $\pi(\beta)$ times (3.8) equal to zero yields

$$\begin{aligned}\Big(\partial_t + \mathbf{v}.\partial_x\Big)\pi(\beta)U_1^\delta &= \pi\,L(\partial_y)\pi U_1^\delta \\ &= -\Big(\partial_T U_0^\delta - \partial_z^{-1}\pi L(\partial_y)QL(\partial_y)U_0^\delta + \pi\,\chi^\delta\Phi_J(U_0^\delta)\Big).\end{aligned} \tag{3.9}$$

Thanks to (3.5) and (3.7), the right-hand side is constant along the integral curves of $\partial_t + \mathbf{v}.\partial_x$. Therefore (3.9) implies that U_1^δ grows linearly along these straight lines unless the constant value is zero. As pointed out in the paragraph before section 3.1, such growth is unacceptable. Thus we must have

$$(\partial_t + \mathbf{v}.\partial_x)\,\pi(\beta)U_1^\delta = 0 \tag{3.10}$$

and

$$\partial_T\partial_z U_0^\delta - \pi L(\partial_y)QL(\partial_y)U_0^\delta + \pi\,\chi^\delta\partial_z\Phi_J(U_0^\delta) = 0. \tag{3.11}$$

As shown in [11], at smooth points of the characteristic variety of $L(\partial_y)$,

$$\pi(\beta)\,L(\partial_y)\,Q\,(\beta)\,L(\partial_y)\,\pi(\beta) \;=\; -P(\partial_x)\,\pi(\beta), \tag{3.12}$$

where the second order differential operator $P(\partial_x)$ is defined in (1.4). Using (3.12) in (3.11), one gets the fundamental equation

$$\partial_T\partial_z U_0^\delta + \,P(\partial_x)U_0^\delta + \pi\,\chi^\delta\partial_z\Phi_J(U_0^\delta) = 0. \tag{3.13}$$

Note in passing that formulas (3.5), (3.7), and (3.10) imply that U_1^δ satisfies the transport equation

$$\partial_t U_1^\delta + \mathbf{v}.\partial_x U_1^\delta = 0. \tag{3.14}$$

It remains to annihilate the product of $Q(\beta)$ with the ε^{p+1} term. This yields

$$(I-\pi(\beta))\partial_z U_2^\delta = -Q(\beta)\Big(\partial_T U_0^\delta + L(\partial_y)\,U_1^\delta + \chi^\delta \Phi_J(U_0^\delta)\Big). \tag{3.15}$$

Thanks to the polarization (3.3), $Q(\beta)\partial_T U_0^\delta = 0$. When $\delta > 0$, (3.15) is solvable and yields

$$(I-\pi(\beta))\,U_2^\delta = -Q(\beta)\,\partial_z^{-1}\Big(L(\partial_y)\,U_1^\delta + \chi^\delta \Phi_J(U_0^\delta)\Big). \tag{3.16}$$

The above calculations pose no constraints on $\pi(\beta)U_2^\delta$ and $\pi(\beta)U_1^\delta|_{t=0}$. For simplicity we set them equal to zero, and using (3.14) we find that

$$\pi(\beta)U_1^\delta = \pi(\beta)U_2^\delta = 0. \tag{3.17}$$

With these choices, (3.16) implies that

$$(\partial_t + \mathbf{v}.\partial_x)\; U_j^\delta = 0 \qquad \text{for} \quad j = 0,1,2. \tag{3.18}$$

Corrector summary. Once U_0^δ is known with the Fourier transform vanishing on a neighborhood of $\zeta = 0$, the correctors are defined by

$$U_1^\delta = -\partial_z^{-1}\,Q(\beta)\,L(\partial_y)\,U_0^\delta, \tag{3.19}$$

$$U_2^\delta = -\partial_z^{-1}\,Q(\beta)\,\Big(L(\partial_y)\,U_1^\delta + \chi^\delta \Phi_J(U_0^\delta)\Big). \tag{3.20}$$

Residual summary. With the profiles defined in this way,

$$\begin{aligned} &L\Big((\varepsilon\partial_T,0)+\partial_y+\frac{\beta}{\varepsilon}\partial_z\Big)\varepsilon^p U^\delta + \chi^\delta(D_z)\,\Phi(\varepsilon^p U^\delta) \\ &\qquad = \varepsilon^{p+2}\Big(\varepsilon\partial_T U_2^\delta + L(\partial_y)U_2^\delta + \partial_T U_1^\delta\Big) + \chi^\delta(D_z)\Big[\Phi(\varepsilon^p U^\delta) - \Phi_J(\varepsilon^p U_0^\delta)\Big]. \end{aligned} \tag{3.21}$$

4. Solvability of the profile equations for $\mathbf{U_0}(\boldsymbol{T},\boldsymbol{x},\boldsymbol{z})$. U_0^δ must be constructed satisfying the equations

$$\begin{aligned} &\pi(\beta)\,U_0^\delta = U_0^\delta, \\ &\partial_t U_0^\delta + \mathbf{v}.\partial_x U_0^\delta = 0, \\ &\partial_T\partial_z U_0^\delta + P(\partial_x)U_0^\delta + \pi(\beta)\,\chi^\delta(D_z)\,\partial_z\Phi_J(U_0^\delta) = 0. \end{aligned} \tag{4.1}$$

Equations (2.7) satisfied by U_0 are obtained by setting $\delta = 0$. Taking advantage of the middle equations of (2.7) and (4.1), define $\mathrm{U}_0(T,x,z)$ and $\mathrm{U}_j^\delta(T,x,z)$ by

$$U_0(T,t,x,z) := \mathrm{U}_0(T,x-\mathbf{v}t,z), \qquad U_j^\delta(T,t,x,z) := \mathrm{U}_j^\delta(T,x-\mathbf{v}t,z).$$

The last equation in (2.7) is then equivalent to

$$\partial_T\,\partial_z \mathrm{U}_0 + P(\partial_x)\,\mathrm{U}_0 + \pi(\beta)\,\partial_z\Phi_J(\mathrm{U}_0) = 0. \tag{4.2}$$

The *nonlinear diffractive pulse equation* (4.2) has $T = 0$ as a characteristic surface. In the wave train case one would have found an equation of nonlinear Schrödinger type at this stage. It too has $T = 0$ characteristic. In both cases the equation gives rise to a well-defined time evolution, at least locally in T.

Before proving this, we first prove that while the diffractive pulse equation appears to have $d+2$ independent variables (t,x,z), it actually involves one less direction of differentiation.

PROPOSITION 4.1. *The matrix $\partial^2\tau/\partial\xi_j\partial\xi_k$ has rank at most $d-1$. In fact $\xi_0 \in \ker \partial^2\tau(\xi_0)/\partial\xi_j\partial\xi_k$.*

Proof. Since $\tau(\xi)$ is homogeneous of degree 1, it follows that for all j, $\partial\tau(\xi)/\partial\xi_j$ is homogeneous of degree zero. Therefore

$$\frac{d}{d\lambda}\frac{\partial\tau(\lambda\xi)}{\partial\xi_j} \;=\; 0.$$

Expanding the left-hand side using the chain rule yields

$$\sum_i \xi_i \,\frac{\partial^2\tau(\lambda\xi)}{\partial\xi_i\partial\xi_j} \;=\; 0.$$

Setting $\lambda = 1$ yields the desired result. ∎

In coordinates so that $\xi_0 = (1,0,\ldots,0)$ this implies that $\tau_{1,j} = \tau_{j,1} = 0$ so that

$$P(\partial_x) = -\frac{1}{2}\sum_{j,k=\mathbf{2}}^{d} \frac{\partial^2\tau(1,0,\ldots,0)}{\partial\xi_j\partial\xi_k}\,\frac{\partial^2}{\partial x_j\partial x_k}.$$

This decrease in the number of spatial dimensions decreases the complexity of the numerical implementation of the results of this paper.

The local solvability of the nonlinear diffractive pulse equation can be proved in the Sobolev spaces $H^s(\mathbb{R}^{1+d})$ with $s > (d+1)/2$ by standard methods. The results would apply for general nonlinearities. A weakness is that for the Fourier transform of $\mathrm{U}_0(T,x,z)$ with respect to the x,z variables one has only L^2 control locally.

The nonlinear term in the profile equation is always a polynomial. Because of this, we have the luxury of working in spaces related to the Wiener algebra which give us L^1 control of the Fourier transform. That in turn permits us to get L^∞ control of $\mathcal{F}_z U_0$ by estimates entirely on the Fourier side. These L^∞ estimates imply that $(I-\chi^\delta)\mathrm{U}_0 = O(\delta)$ as $\delta \to 0$. The usual strategy to obtain sup norm estimates for Fourier transforms is to prove decay rates as $x, z \to \infty$. The argument completely on the Fourier side circumvents that avenue. We do not prove any decay rates beyond those implied by being in $\cap_s H^s(\mathbb{R}^{d+1}_{x,z})$.

DEFINITION 4.2. *The* Wiener algebra $\mathbb{A}(\mathbb{R}^M)$ *is the Banach space of tempered distributions on $\mathbb{R}^M$ with the property that their Fourier transform belongs to $L^1(\mathbb{R}^M)$. The norm is the L^1 norm of the Fourier transform.*

Recall that for any $1 \le p \le \infty$ the map

$$L^1 \times L^p \ni f, g \quad\to\quad f * g \in L^p$$

is a continuous bilinear map from $L^1 \times L^p$ to L^p and

$$\|f * g\|_{L^p} \le \|f\|_{L^1}\,\|g\|_{L^p}. \tag{4.3}$$

The inequality (4.3) with $p = 1$ shows that the map $\mathrm{U} \to \Phi_J(\mathrm{U})$ maps bounded sets of $\mathbb{A}$ to bounded sets of $\mathbb{A}$. To study the continuity of the map note that the difference $\Phi_J(\mathrm{U}) - \Phi_J(\mathrm{V})$ can be expressed as

$$\Phi_J(\mathrm{U}) - \Phi_J(\mathrm{V}) = \sum_i^d P_i(\mathrm{U},\mathrm{V})\left(\mathrm{U}_i - \mathrm{V}_i\right)$$

with polynomials P_i of degree less than J. Then inequality (4.3) shows that the map $\mathrm{U} \to \Phi_J(\mathrm{U})$ is uniformly Lipschitzian on bounded subsets of $\mathbb{A}$ to $\mathbb{A}$.

The initial value problem for the profile equation (4.2) with initial value

$$\mathrm{U}|_{T=0} = G(x,z)$$

is equivalent to the integral identities

$$\widehat{\mathrm{U}}(T) = e^{-iP(\xi)T/\zeta}\,\widehat{G} + \int_0^T e^{-iP(\xi)(T-\sigma)/\zeta}\;\pi(\beta)\,\mathcal{F}\big(\Phi_J(\mathrm{U}(\sigma))\big)\;d\sigma \tag{4.4}$$

for $0 \le T \le \underline{T}$.

The multipliers $e^{iP(\xi)t/\zeta}$ have modulus one so they define isometries on $\mathbb{A}(\mathbb{R}^{d+1}_{x,z})$. This, together with the uniform Lipschitzian property, is enough to make Picard's classical existence proof work, yielding the following result.

PROPOSITION 4.3. *For each $G \in \mathbb{A}$ there is a $T_* = T_*(G) \in \,]0,\infty]$ and a unique maximal solution $\mathrm{U} \in C\big([0,T_*[\,;\mathbb{A}\big)$ to the profile equation (4.2), which in addition satisfies the initial condition $\mathrm{U}|_{T=0} = G$. The time T_* is uniformly strictly positive on bounded subsets of $\mathbb{A}$, and if $T_* < \infty$, then*

$$\lim_{T\to T_*} \|\mathrm{U}(T)\|_{\mathbb{A}} \;=\; \infty. \tag{4.5}$$

The next result is a regularity theorem which asserts that if the initial data lies in a smaller Banach space $\mathbb{B}$, then the maximal solution is a continuous function with values in $\mathbb{B}$.

DEFINITION 4.4. *A Banach space $\mathbb{B} \subset \mathbb{A}$ is* admissible *if it has the following three properties:*

1. *The inclusion map $\mathbb{B} \to \mathbb{A}$ is continuous.*
2. *The map $\mathrm{U} \to \Phi_J(\mathrm{U})$ maps $\mathbb{B}$ to itself and is uniformly Lipschitzian on subsets of $\mathbb{B}$ which are bounded in $\mathbb{A}$.*
3. *For $T \neq 0$, the Fourier multipliers $e^{iTP(\xi)/\zeta}$ are isometries from $\mathbb{B}$ to itself.*

The following are examples of admissible Banach spaces.

Example 4.5. If $1 < p \le \infty$, then $\mathbb{B} := \{\mathrm{U} \in \mathbb{A} : \widehat{\mathrm{U}} \in L^p\}$ is admissible.

Example 4.6. If $s > (d+1)/2$, then $H^s(\mathbb{R}^{d+1})$ is admissible.

Example 4.7. If $\mathbb{B}_1$ and $\mathbb{B}_2$ are admissible, then so is the intersection $\mathbb{B}_1 \cap \mathbb{B}_2$.

Proof for Example 4.5. Only property 2 in the definition is not immediate. One needs to prove that for every $R > 0$ there is a constant C so that if $\|\mathrm{U}\|_{\mathbb{A}} \le R$ and $\|\mathrm{V}\|_{\mathbb{A}} \le R$, then

$$\big\|\Phi_J(\mathrm{U}) - \Phi_J(\mathrm{V})\big\|_{\mathbb{B}} \le C\,\big\|\mathrm{U}-\mathrm{V}\big\|_{\mathbb{B}}. \tag{4.6}$$

Taylor's theorem implies that

$$\Phi_J(\mathrm{U}) - \Phi_J(\mathrm{V}) = \Psi(\mathrm{U},\mathrm{V})(\mathrm{U}-\mathrm{V}),$$

where Ψ is a matrix-valued homogeneous polynomial of degree $J-1$.

To estimate the L^p norm of the Fourier transform use Young's inequality

$$\begin{aligned}\big\|\mathcal{F}\Big(\Phi_J(\mathrm{U}) - \Phi_J(\mathrm{V})\Big)\big\|_{L^p} &\le \big\|\mathcal{F}\Big(\Psi(\mathrm{U},\mathrm{V})\Big)\big\|_{L^1}\big\|\mathcal{F}(\mathrm{U}-\mathrm{V})\big\|_{L^p}\\ &\le C(R)\,\|\mathrm{U}-\mathrm{V}\|_{\mathbb{B}}. \qquad \square\end{aligned}$$

PROPOSITION 4.8. *If $\mathbb{B}$ is admissible and $G \in \mathbb{B}$, then the maximal solution found in Proposition* 4.3 *satisfies*

$$\mathrm{U} \in C([0, T_*[\,; \mathbb{B}). \tag{4.7}$$

Proof. From the admissibility properties one easily demonstrates using Picard's method that the integral equation (4.4) has a maximal solution

$$\mathrm{U} \in C([0, T^*(G)[\,; \mathbb{B}),$$

and if $T^* < \infty$, then

$$\lim_{T \to T^*} \|\mathrm{U}(T)\|_{\mathbb{B}} \;=\; \infty. \tag{4.8}$$

Since this solution is continuous with values in $\mathbb{A}$ it follows that $T^*(G) \le T_*(G)$. The result of the proposition follows from establishing the inequality $T^*(G) \ge T_*(G)$.

The proof is indirect. We suppose on the contrary that $T^*(G) < T_*(G)$ and derive a contradiction.

If $T^*(G) < T_*(G)$, then U is continuous on $[0, T^*]$ with values in $\mathbb{A}$, and so there is an $R < \infty$ so that

$$\|\mathrm{U}(T)\|_{\mathbb{A}} \le R \qquad \text{for} \quad 0 \le T \le T^*.$$

Taking the $\mathbb{B}$ norm of (4.4) and using the last two properties from the definition of admissibility yields

$$\|\mathrm{U}(T)\|_B \le \|G\|_B + \int_0^T C\,\|\mathrm{U}(\sigma)\|_{\mathbb{B}}\, d\sigma \qquad \text{for} \quad 0 \le T < T^*.$$

Gronwall's inequality implies that

$$\|\mathrm{U}(T)\|_B \le \|G\|_B\; e^{CT} \qquad \text{for} \quad 0 \le T < T^*.$$

In particular, (4.8) is violated. This contradiction proves the proposition. ∎

Our main existence result is a corollary of Propositions 4.9 and 4.10 below.

PROPOSITION 4.9. *Define $\mathbb{B}$ to be the closed subspace of $\mathbb{A}(\mathbb{R}^{d+1}_{x,z})$ consisting of functions* U *such that* (i) $\widehat{\mathrm{U}} \in L^\infty(\mathbb{R}^{d+1}_{\xi,\zeta})$, *and* (ii) *for all $\mu > 0$, $\widehat{\mathrm{U}}$ is uniformly continuous on $\{|\zeta| \ge \mu\}$. Then $\mathbb{B}$ is admissible.*

Proof. $\mathbb{B}$ is a closed subspace of Example 4.5 with $p = \infty$. Thus to prove that $\mathbb{B}$ is admissible it suffices to show that Φ_J maps $\mathbb{B}$ to itself. In fact, more is true. If $\widehat{\mathrm{U}} \in L^1 \cap L^\infty$ (which is true if $\mathrm{U} \in \mathbb{B}$), then $\mathcal{F}_{x,z}\big(\Phi_J(\mathrm{U})\big)$ is bounded and uniformly continuous, which implies that $\Phi_J(\mathrm{U}) \in \mathbb{B}$.

To prove the stronger assertion of the last sentence, write Φ_J as a sum of terms each of which is a product of a monomial of order $J-1$ and a monomial of order 1. The Fourier transform of the first factor belongs to L^1 and the Fourier transform of the second belongs to L^∞. The desired result follows from the fact that the convolution of an element of L^1 with an element of L^∞ is uniformly continuous. ∎

PROPOSITION 4.10. *If $0 \le m \in \mathbb{Z}$ and $1 \le p \le \infty$, then the subspace*

$$\mathbb{B}^{m,p} \;:=\; \Big\{ \mathrm{U} \in \mathbb{A} \,:\, \langle \eta \rangle^m\, \widehat{\mathrm{U}}(\eta) \,\in L^p(\mathbb{R}^N_\eta) \Big\}$$

is admissible.

Proof. This proof follows [23]. The only sticky point is estimate (4.6). Toward that end one must show that for K-fold products one has

$$\text{(4.9)}\quad \left\|\mathrm{U}_j\right\|_{\mathbb{A}} \;\le\; R \quad\Rightarrow\quad \left\|\langle\eta\rangle^m\,\widehat{\mathrm{U}}_1 * \widehat{\mathrm{U}}_2 * \cdots * \widehat{\mathrm{U}}_K\right\|_{L^p} \;\le\; C(K,R)\sum_j \left\|\langle\eta\rangle^m\,\widehat{\mathrm{U}}_j\right\|_{L^p}.$$

Written out, this becomes

$$\text{(4.10)}\quad \begin{aligned}&\Big\|\int \langle\eta\rangle^m\,\widehat{\mathrm{U}}_1(\eta-\eta_1)\,\widehat{\mathrm{U}}_2(\eta_1-\eta_2)\cdots\widehat{\mathrm{U}}_K(\eta_K)\;\,d\eta_1\,d\eta_2\cdots d\eta_{K-1}\,d\eta_K\Big\|_{L^p}\\ &\qquad\le\; C(K,R)\sum_j\left\|\langle\eta\rangle^m\widehat{\mathrm{U}}_j\right\|_{L^p}.\end{aligned}$$

Note that

$$\eta = (\eta-\eta_1)+(\eta_1-\eta_2)+\cdots+(\eta_{K-1}-\eta_K)+\eta_K,$$

so

$$|\eta| \le |\eta-\eta_1|+|\eta_1-\eta_2|+\cdots+|\eta_{K-1}-\eta_K|+|\eta_K|.$$

Define $\eta_0 := \eta, \eta_{K+1} := 0$, so the summands on the right are equal to $|\eta_{j-1}-\eta_j|$ for $1\le j\le K+1$. The integral in (4.10) is split into $K+1$ pieces, the integrals over sets

$$E(j) \;:=\; \Big\{(\eta,\eta_1,\ldots,\eta_K)\;:\;|\eta_{j-1}-\eta_j| = \max_{1\le k\le K}\big\{|\eta_{k-1}-\eta_k|\big\}\Big\}.$$

The sets E_j overlap in measure zero sets, so it suffices to show that

$$\begin{aligned}&\Big\|\int_{E(j)}\langle\eta\rangle^m\,\widehat{\mathrm{U}}_1(\eta-\eta_1)\,\widehat{\mathrm{U}}_2(\eta_1-\eta_2)\cdots\widehat{\mathrm{U}}_K(\eta_K)\;\,d\eta_1\,d\eta_2\cdots d\eta_{K-1}\,d\eta_K\Big\|_{L^P}\\ &\qquad\le\; C(K,R)\sum_j\left\|\langle\eta\rangle^m\widehat{\mathrm{U}}_j\right\|_{L^p}.\end{aligned}$$

On $E(j)$, $|\eta|\le K|\eta_{j-1}-\eta_j|$ so

$$\begin{aligned}&\int_{E(j)}\langle\eta\rangle^m\,|\widehat{\mathrm{U}}_1(\eta-\eta_1)|\cdot|\widehat{\mathrm{U}}_2(\eta_1-\eta_2)|\cdots|\widehat{\mathrm{U}}_K(\eta_K)|\;d\eta_1 d\eta_2\cdots d\eta_{K-1}d\eta_K\\ &\quad\le C(K)\int_{E(j)}\langle\eta_{j-1}-\eta_j\rangle^m\,|\widehat{\mathrm{U}}_1(\eta-\eta_1)|\cdot|\widehat{\mathrm{U}}_2(\eta_1-\eta_2)|\\ &\qquad\qquad\cdots|\widehat{\mathrm{U}}_K(\eta_K)|\;d\eta_1 d\eta_2\cdots d\eta_{K-1}d\eta_K.\end{aligned}$$

Young's inequality bounds the L^p norm of the integral on the right by

$$\left\|\langle\eta\rangle^m\widehat{\mathrm{U}}_j\right\|_{L^p}\prod_{k\ne j}\left\|\widehat{\mathrm{U}}_k\right\|_{L^1} \;=\; C(R)\left\|\langle\eta\rangle^m\widehat{\mathrm{U}}_j\right\|_{L^p}.$$

This completes the proof. ∎

DEFINITION 4.11. *$\mathcal{B}$ is the Fréchet space of tempered distributions $\mathrm{V}(x,z)$ so that*

1. *$\widehat{\mathrm{V}}(\xi,\zeta)\in L^\infty(\mathbb{R}^{d+1})$ and for every $\mu>0$ is uniformly continuous on the set $\{(\xi,\zeta)\;:\;|\zeta|\ge\mu\}$.*

2. *For every nonnegative integer* m

$$\langle \xi, \zeta \rangle^m \, \widehat{\mathrm{V}}(\xi, \zeta) \in L^\infty(\mathbb{R}^{d+1}).$$

Combining Propositions 4.9 and 4.10 shows that $\mathcal{B}$ is the intersection of admissible spaces. By Example 4.7, this implies the following corollary.

COROLLARY 4.12. *Suppose that* $\mathrm{U} \in C\big([0, T_*[\,; \mathbb{A})\big)$ *is a maximal solution of the profile equation* (4.2) *and that* $\mathrm{U}\big|_{T=0} \in \mathcal{B}$. *Then*

$$\mathrm{U} \in C\big([0, T_*[\,; \mathcal{B}\big). \tag{4.11}$$

This corollary gives us more than enough control on the leading profile to carry out our analysis. The most interesting aspect is the sup norm control near $\zeta = 0$ *without* continuity.

5. Construction of $\mathbf{U_0^\delta}$. A perturbation argument is the key to solving the $\delta > 0$ equations (4.1). The third equation in (4.1) is equivalent to

$$\partial_T \partial_z \mathrm{U}_0^\delta + P(\partial_x)\mathrm{U}_0^\delta + \pi(\beta)\, \chi^\delta(D_z)\, \partial_z \Phi_J(\mathrm{U}_0^\delta) = 0. \tag{5.1}$$

Multiplying (4.2) by $\chi^\delta(D_z)$ yields

$$\partial_T\, \partial_z \chi^\delta(D_z)\, \mathrm{U}_0 + P(\partial_x)\, \chi^\delta(D_z)\, \mathrm{U}_0 + \pi(\beta)\, \chi^\delta(D_z)\, \partial_z \Phi_J(\mathrm{U}_0) = 0. \tag{5.2}$$

This equation resembles (5.1), as demonstrated in the next proposition, which shows that the solution of (5.1) can be obtained as a small perturbation of $\chi^\delta(D_z)\, \mathrm{U}_0$.

PROPOSITION 5.1. *Suppose that* $\mathrm{U}_0 = \pi(\beta)\mathrm{U}_0 \in C\big([0, \underline{T}]\,; \cap_m \mathbb{B}^{m,\infty}\big)$ *satisfies* (4.2). *Then there is a* $\delta_0 > 0$ *so that for* $0 < \delta < \delta_0$ *the initial value problem defined by* (5.2) *with initial condition*

$$\mathrm{U}_0^\delta\big|_{T=0} = \chi^\delta(D_z)\mathrm{U}_0\big|_{T=0} \tag{5.3}$$

has a unique solution $\mathrm{U}_0^\delta \in C([0, \underline{T}]\,; \cap_m \mathbb{B}^{m,\infty})$, *and for all* $1 \le q < \infty$ *and* $0 \le m < \infty$,

$$\sup_{0 \le T \le \underline{T}} \big\| \mathrm{U}_0^\delta(T) - \chi^\delta\, \mathrm{U}_0(T) \big\|_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1}_{x,z})} \;=\; O(\delta^{1/q}).$$

Furthermore the Fourier transform $\mathcal{F}_z\big(\mathrm{U}_0^\delta\big)$ *vanishes identically on* $|\zeta| \le \delta$.

Proof of Proposition 5.1. Begin with the proof that U_0^δ has a Fourier transform with respect to z vanishing on $|\zeta| \le \delta$. For any $\gamma(\zeta) \in C_0^\infty(\mathbb{R})$ supported on $|\zeta| \le 1$ define $\gamma^\delta(\zeta) := \gamma(\zeta/\delta)$. It suffices to show that $\mathcal{F}_z \gamma^\delta(D_z)\mathrm{U}_0^\delta(T) = 0$.

The choice of γ implies that $\gamma^\delta \chi^\delta = 0$. Thus multiplying (5.1) by $\gamma^\delta(D_z)$ annihilates the nonlinear term. This implies that $\gamma^\delta(D_z)\mathrm{U}_0^\delta \in C\big([0, T_*[\,; \cap_s H^s\big)$ satisfies

$$\Big(\partial_T \partial_z + P(\partial_x)\Big)\, \gamma^\delta(D_z)\mathrm{U}_0^\delta \;=\; 0. \tag{5.4}$$

In addition $\gamma^\delta(D_z)\mathrm{U}_0^\delta$ vanishes when $T = 0$.

It follows from the basic H^s conservation law for the linear diffractive pulse equation [4] that for all s and all $T \in [0, T_*[$,

$$\|\gamma^\delta(D_z)\, \mathrm{U}_0^\delta(T)\|_{H^s(\mathbb{R}^{d+1}_{x,z})} = \|\gamma^\delta(D_z)\, \mathrm{U}_0^\delta(0)\|_{H^s(\mathbb{R}^{d+1}_{x,z})} = 0. \tag{5.5}$$

Identity (5.5) implies the second assertion of the proposition.

The strategy for proving the $O(\delta^{1/q})$ estimate in the proposition is to construct U_0^δ as a perturbation of $\chi^\delta(D_z)\,\mathrm{U}_0$. Define the perturbation W^δ by

$$\mathrm{W}^\delta := \mathrm{U}_0^\delta - \chi^\delta(D_z)\mathrm{U}_0\,. \tag{5.6}$$

Subtract (5.1) from (5.2) to show that U_0^δ is a solution if and only if W^δ satisfies the initial value problem

$$\begin{aligned} &\partial_T\,\partial_z \mathrm{W}^\delta + P(\partial_x)\,\mathrm{W}^\delta + \pi(\beta)\,\chi^\delta(D_z)\,\partial_z\left(\Phi_J(\mathrm{U}_0^\delta) - \Phi_J(\mathrm{U}_0)\right) = 0,\\ &\mathrm{W}^\delta\big|_{T=0} = 0. \end{aligned} \tag{5.7}$$

Note that

$$\Phi_J(\mathrm{U}_0^\delta) - \Phi_J(\mathrm{U}_0) = \Phi_J(\chi^\delta\mathrm{U}_0 + \mathrm{W}^\delta) - \Phi_J(\chi^\delta\mathrm{U}_0 + (I-\chi^\delta)\mathrm{U}_0).$$

Since $\chi^\delta\mathrm{U}_0(T)$ is bounded in the admissible subspace $\mathbb{B}^{m,\infty}$ uniformly for all $0 < \delta \le 1$ and $0 \le t \le \underline{T}$, the second condition in the definition of admissibility implies that as long as $\|\mathrm{W}^\delta\|_{\mathbb{B}^{m,\infty}} \le 1$,

$$\begin{aligned} \left\|\Phi_J(\mathrm{U}_0^\delta) - \Phi_J(\mathrm{U}_0)\right\|_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1})} \le C(m,q)\,\Big(&\left\|\mathrm{W}^\delta(T)\right\|_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1})}\\ &+ \left\|(I-\chi^\delta)\mathrm{U}_0(T)\right\|_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1})}\Big)\,. \end{aligned} \tag{5.8}$$

Fix $m \ge 0$ and $1 \le q \le \infty$. If $\|\mathrm{W}^\delta(T)\|_{\mathbb{B}^{m,q}} \le 1$ for $0 \le T \le \underline{T}$, define $T_* = T_*(m,q,\delta) = \underline{T}$. Otherwise define

$$T_*(m,q,\delta) \;:=\; \inf\{T \in [0,\underline{T}] \;:\; \|\mathrm{W}^\delta(T)\|_{\mathbb{B}^{m,q}} = 1\}.$$

The homogeneous linear diffractive pulse equation generates a unitary group on each $\mathbb{B}^{m,q}$. The inhomogeneous version of this estimate implies that for $T \in [0, T_*]$,

$$\begin{aligned} \|\mathrm{W}^\delta(T)\|_{\mathbb{B}^{m,q}} \le \int_0^T C(m)\,\Big(&\left\|\mathrm{W}^\delta(\sigma)\right\|_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1})}\\ &+ \left\|(I-\chi^\delta)\mathrm{U}_0(\sigma)\right\|_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1})}\Big)\,d\sigma. \end{aligned} \tag{5.9}$$

Gronwall's inequality then shows that

$$\|\mathrm{W}^\delta(T)\|_{\mathbb{B}^{m,q}} \le C(m)\int_0^T \left\|(I-\chi^\delta)\mathrm{U}_0(T)\right\|_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1})}\,dt\;e^{C(m)T}. \tag{5.10}$$

To proceed we need the following lemma.

LEMMA 5.2. *For any $m \ge 0$, $1 \le q < \infty$, and $M > q + (d+1)/q$ there is a constant $C = C(m,q,M)$ so that for all $\delta > 0$ and all $W \in \mathbb{B}^{M,\infty}$,*

$$\left\|(I-\chi^\delta)W\right\|_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1})} \;\le\; C\,\delta^{1/q}\,\|W\|_{\mathbb{B}^{M,\infty}}.$$

Remark. In contrast note that for $W \in \cap_s H^s$, $||(I-\chi^\delta)W\|_{H^s} = o(1)$ as $\delta \to 0$, but there is no rate of convergence.

Proof of Lemma. By definition

$$\left\|(I-\chi^\delta)W\right\|^q_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1})} \;=\; \int \left|1-\chi(\zeta/\delta)\right|^q\,\left|\widehat{W}(\xi,\zeta)\right|^q\,\langle\xi,\zeta\rangle^{mq}\,d\xi\,d\zeta.$$

Use the estimate

$$|\widehat{W}(\xi,\zeta)| \le \langle \xi,\zeta\rangle^{-M}\, \|\widehat{W}\|_{\mathbb{B}^{M,\infty}}$$

to find

$$\big\|(I-\chi^\delta)W\big\|^q_{\mathbb{B}^{m,q}} \le \|\widehat{W}\|^q_{\mathbb{B}^{M,\infty}} \int \big|1-\chi(\zeta/\delta)\big|^q\ \langle \xi,\zeta\rangle^{-Mq}\ \langle \xi,\zeta\rangle^{mq}\ d\xi\, d\zeta.$$

The integral on the right is over $|\zeta| \le \delta$, so for $Mq > mq + d + 1$, the integral is $O(\delta)$, which proves the lemma. □

This lemma implies that

$$\int_0^{\underline{T}} \big\|(I-\chi^\delta)\mathrm{U}_0(T)\big\|_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1})}\ dt\ =\ O(\delta^{1/q}).$$

Thus, one can choose $\delta_0(m,q)$ so that for $0 < \delta < \delta_0$,

$$C(m)\int_0^{\underline{T}} \big\|(I-\chi^\delta)\mathrm{U}_0(T)\big\|_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1})}\ dt\ e^{C(m)\underline{T}} < \frac{1}{2}.$$

Then (5.10) shows that if $T_* < \underline{T}$, then

$$\|\mathrm{W}^\delta(T)\|_{\mathbb{B}^{m,q}} \le \frac{1}{2}$$

for $0 < t < T_*$. Since this contradicts the definition of T_* we conclude that $T_* = \underline{T}$ and that (5.10) holds for $0 < T < \underline{T}$. In particular, the evolution equation for U_0^δ is solvable up to $\underline{T}$.

In addition, estimate (5.10) implies the $O(\delta^{1/q})$ convergence rate for the value m, q fixed at the start. Since m, q is arbitrary, the proof of Proposition 5.1 is complete. □

Combining the convergence results of Proposition 5.1 and Lemma 5.2 yields

$$\sup_{0\le t\le T} \|\mathrm{U}_0(T) - \mathrm{U}_0^\delta(T)\|_{\mathbb{B}^{m,q}(\mathbb{R}^{d+1}_{x,z})}\ =\ O(\delta^{1/q}). \tag{5.11}$$

Note that the smallest upper bound occurs for the case $q = 1$ corresponding to the Wiener algebra.

6. Estimate for the residual. Suppose that $\underline{T}$ is smaller than the maximal existence time for U_0 in the sense that a solution of (4.2) is known in the space $C\big([0,\underline{T}]\,;\cap_m \mathbb{B}^{m,\infty}(\mathbb{R}^{d+1}_{x,z})\big)$. Then Proposition 5.1 shows that for $0 < \delta < \delta_0$, $U_0^\delta(T,t,x,z) = \mathrm{U}_0^\delta(T, x-\mathbf{v}t, z)$ exists on $[0,\underline{T}]$ and $\mathcal{F}_z U_0^\delta$ vanishes on a neighborhood of $\zeta = 0$. For $\delta > 0$, the equations for the correctors U_1^δ, U_2^δ are solvable, so

$$U^\delta(\varepsilon,T,t,x,z) := U_0^\delta + \varepsilon U_1^\delta + \varepsilon^2 U_2^\delta$$

is well defined for $0 \le t \le \underline{T}/\varepsilon$. The residual equation (3.21) is then

$$\begin{aligned} R^\delta(\varepsilon,T,t,x,z) &:= L\Big((\varepsilon\partial_T,0) + \partial_y + \frac{\beta}{\varepsilon}\partial_z\Big)\,\varepsilon^p U^\delta + \chi^\delta(D_z)\,\Phi(\varepsilon^p U^\delta) \\ &= \varepsilon^{p+1}\Big(\varepsilon^2\partial_T U_2^\delta + \varepsilon L(\partial_y)U_2^\delta + \varepsilon\partial_T U_1^\delta\Big) + \chi^\delta(D_z)\Big[\Phi(\varepsilon^p U^\delta) - \Phi_J(\varepsilon^p U_0^\delta)\Big]. \end{aligned} \tag{6.1}$$

This section is devoted to estimates for the right-hand side of (6.1). There are at least two subtle points. The first is that though the residual is formally $O(\varepsilon^{2p+1})$ it involves correctors which blow up as $\delta \to 0$. Care must be taken about small values of δ. The second point is that the residual involves the smooth function Φ, which need not be polynomial. Therefore, the Wiener algebra need not be invariant. The estimates are therefore done in the scale of Sobolev spaces.

PROPOSITION 6.1. *Suppose that U^δ and $\underline{T}$ are as above, δ_0 is as in Proposition 5.1, and $s > (d+1)/2$. Then there is a constant $C = C(s)$ so that for all*

$$0 \le t < \infty \quad \cap \quad 0 \le T \le \underline{T} \quad \cap \quad 0 < \delta \le \delta_0 \quad \cap \quad 0 \le \varepsilon \le \delta, \tag{6.2}$$

one has

$$\|R^\delta(\varepsilon, T, t, x, z)\|_{H^s(\mathbb{R}^{d+1}_{x,z})} \;\le\; C\,\varepsilon^{p+1}\Big(\frac{\varepsilon}{\delta^2} + \varepsilon^p\Big). \tag{6.3}$$

Proof. The first step is to estimate the size of the correctors U_1^δ, U_2^δ. The formulas for these functions involve U_0^δ and, most importantly, the operator ∂_z^{-1}. The formula for U_1^δ involves ∂_z^{-1}, while the formula for U_2^δ involves ∂_z^{-2} since it has a term with ∂_z^{-1} applied to U_1^δ. The application of the operator ∂_z^{-1} introduces a factor $1/\delta$ in estimates since the support of U_j^δ is in $|\zeta| > \delta$. The boundedness of the family U_0^δ in $\mathbb{B}^{m,\infty}$ yields the following estimates for the correctors for $0 < \delta \le \delta_0$:

$$\forall m\,, \forall\, 0 \le t < \infty, \quad \sup_{0 \le T \le \underline{T}} \; \big\|U_j^\delta(T, t, x, z)\big\|_{\mathbb{B}^{m,\infty}(\mathbb{R}^{d+1}_{x,z})} \le \frac{C(m)}{\delta j}. \tag{6.4}$$

Inserted in the definition of U^δ, this estimate proves that for $t \in \mathbb{R}$ and $0 \le T \le \underline{T}$,

$$\big\|U^\delta(T,t)\big\|_{\mathbb{B}^{m,\infty}(\mathbb{R}^{d+1}_{x,z})} \le C(m)\Big(1 + \frac{\varepsilon}{\delta} + \frac{\varepsilon^2}{\delta^2}\Big).$$

Thus $U^\delta(T,t)$ is uniformly bounded in $\mathbb{B}^{m,\infty}$ for the parameter range (6.2). This is a key element in the estimate of the second term on the right-hand side of (6.1).

For the right-hand side of (6.1) we also need an estimate for $\varepsilon\partial_T U_1^\delta$ and $\varepsilon^2\partial_T U_2^\delta$. Start with an estimate for the T derivative of U_0^δ. The evolution equation (5.1) yields

$$\partial_T \mathrm{U}_0^\delta \;=\; -\partial_z^{-1} P(\partial_x)\mathrm{U}_0^\delta - \pi(\beta)\,\chi^\delta(D_z)\,\Phi_J(\mathrm{U}_0^\delta).$$

Together with the uniform boundedness of U_0^δ this yields

$$\|\partial_T U_0^\delta\|_{\mathbb{B}^{m,\infty}} \;\le\; \frac{C(m)}{\delta}. \tag{6.5}$$

The factor $1/\delta$ comes from the norm of ∂_z^{-1} acting on functions with spectrum in $|\zeta| \ge \delta$. Differentiating (3.19) and (3.20) with respect to T yields

$$\partial_T U_1^\delta = -\partial_z^{-1} Q(\beta) L(\partial_y)\partial_T U_0^\delta$$

and

$$\partial_T U_2^\delta = -\partial_z^{-1} Q(\beta)\Big(L(\partial_y)\partial_T U_1^\delta + \chi^\delta(D_z)\pi(\beta)\Phi_J'(U_0^\delta)\partial_T U_0^\delta\Big).$$

Using estimate (6.5) in the equations above yields in turn

$$\|\partial_T U_1^\delta\|_{\mathbb{B}^{m,\infty}} \;\le\; \frac{C(m)}{\delta^2} \qquad \text{and} \qquad \|\partial_T U_2^\delta\|_{\mathbb{B}^{m,\infty}} \;\le\; \frac{C(m)}{\delta^3}.$$

Inserting these estimates into the first term on the right-hand side of (6.1) and using $\varepsilon \le \delta$ yields

$$\left\|\varepsilon^2\partial_T U_2^\delta + \varepsilon L(\partial_y)U_2^\delta + \varepsilon\partial_T U_1^\delta\right\|_{\mathbb{B}^{m,\infty}} \le C(m)\Big(\frac{\varepsilon^2}{\delta^3} + \frac{\varepsilon}{\delta^2} + \frac{\varepsilon}{\delta^2}\Big) \;\le\; \frac{C(m)\,\varepsilon}{\delta^2}. \tag{6.6}$$

Next turn to the second term on the right of (6.1). Taylor's theorem with remainder implies that there are smooth functions G_α so that

$$\Phi(u) \;=\; \sum_{|\alpha|=J} u^\alpha\, G_\alpha(u). \tag{6.7}$$

It follows that

$$\Phi(\lambda u) = \lambda^J \Psi(\lambda, u), \tag{6.8}$$

with smooth Ψ satisfying $\Psi(\lambda, 0) = 0$. An entirely analogous argument shows that there is a $\Xi(\lambda, u)$ vanishing when $u = 0$ so that

$$\Phi(\lambda u) - \Phi_J(\lambda u) = \lambda^{J+1}\,\Xi(\lambda, u). \tag{6.9}$$

Split the nonlinearity on the right-hand side of (6.1),

$$\Phi(\varepsilon^p U^\delta) - \Phi_J(\varepsilon^p U_0^\delta) \;=\; \big[\Phi(\varepsilon^p U^\delta) - \Phi(\varepsilon^p U_0^\delta)\big] + \big[\Phi(\varepsilon^p U_0^\delta) - \Phi_J(\varepsilon^p U_0^\delta)\big].$$

Using (6.8), (6.9), and the fact that $\varepsilon^{pJ} = \varepsilon^{p+1}$ yields

$$\Phi(\varepsilon^p U^\delta) - \Phi_J(\varepsilon^p U_0^\delta) \;=\; \varepsilon^{p+1}\big[\Psi(\varepsilon^p, U^\delta) - \Psi(\varepsilon^p, U_0^\delta)\big] + \varepsilon^{2p+1}\,\Xi(\varepsilon^p, U_0^\delta). \tag{6.10}$$

Since both U^δ and U_0^δ are H^s uniformly bounded, (6.10) and Schauder's lemma imply that for $s > (d+1)/2$,

$$\big\|\Phi(\varepsilon^p U^\delta) - \Phi_J(\varepsilon^p U_0^\delta)\big\|_{H^s(\mathbb{R}^{d+1}_{x,z})} \;\le\; C(s)\,\varepsilon^{p+1}\Big(\|U^\delta - U_0^\delta\|_{H^s(\mathbb{R}^{d+1}_{x,z})} + \varepsilon^p\Big). \tag{6.11}$$

Using (6.4) and $\varepsilon \le \delta$ yields

$$\|U^\delta - U_0^\delta\|_{H^s(\mathbb{R}^{d+1}_{x,z})} \;\le\; C(s)\Big(\frac{\varepsilon}{\delta} + \frac{\varepsilon^2}{\delta^2}\Big) \;\le\; \frac{C(s)\,\varepsilon}{\delta}. \tag{6.12}$$

Combining (6.6), (6.11), and (6.12) yields the estimate (6.3). ∎

7. Proof of (2.12). In the next proposition we prove the third and last convergence result needed to establish our main result.

PROPOSITION 7.1. *Suppose that $\underline{T}$ is smaller than the maximal existence time for U_0 and that U^δ, U^δ are as in the first paragraph of section* 6*, and hence in the space satisfying* (1.3)*. Then there is a positive ε_1 so that for $0 < \varepsilon < \varepsilon_1$, the differential equation* (2.10) *has a unique solution $\mathcal{V}(\varepsilon, t, x, \phi) \in C\big([0, \underline{T}/\varepsilon\,]\,; \cap_s H^s(\mathbb{R}^{d+1})\big)$ satisfying the initial condition*

$$\mathcal{V}(\varepsilon, 0, x, \phi) = U_0(0, 0, x, \phi).$$

In addition, for all s, for $\delta = \varepsilon^{2/5}$, and for $0 < \varepsilon < \varepsilon_1$,

$$\sup_{0\le t\le \underline{T}/\varepsilon} \left\|U^\delta\Big(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big) - \mathcal{V}(\varepsilon, t, x, \phi)\right\|_{H^s(\mathbb{R}^d_x \times \mathbb{R}_\phi)} \le C(s)\varepsilon^{\max\{1/5\,,\,p\}}. \tag{7.1}$$

Proof. The construction of U^δ guarantees that $U^\delta(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi)$ satisfies

$$L\Big(\partial_t, \partial_x + \frac{\xi_0}{\varepsilon}\partial_\phi\Big)\varepsilon^p U^\delta + \Phi(\varepsilon^p U^\delta) = R^\delta\Big(\varepsilon, \varepsilon t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big) + \big(I - \chi^\delta(D_z)\big)\,\Phi(\varepsilon^p U^\delta). \tag{7.2}$$

The strategy is to construct $\mathcal{V}$ as a perturbation, $\mathcal{E}^\delta$, of U^δ. Define

$$\mathcal{E}^\delta(\varepsilon, t, x, \phi) \;:=\; \mathcal{V}(\varepsilon, t, x, \phi) - U^\delta\Big(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big). \tag{7.3}$$

The equation for $\mathcal{V}$ is rewritten as an equation for $\mathcal{E}^\delta$. The equation for $\mathcal{E}^\delta$ is then analyzed to show that the perturbation remains small for $0 \le t \le \underline{T}/\varepsilon$.

Subtracting (2.10) from (7.2) yields

$$\begin{aligned} L\Big(\partial_t, \partial_x + \frac{\xi_0}{\varepsilon}\partial_\phi\Big)\varepsilon^p \mathcal{E}^\delta = &-R^\delta\Big(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big) \\ &- \big(I - \chi^\delta(D_z)\big)\Phi\Big(\varepsilon^p U^\delta\Big(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big)\Big) \\ &+ \Big[\Phi\Big(\varepsilon^p U^\delta\Big(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big)\Big) - \Phi(\varepsilon^p \mathcal{V})\Big]. \end{aligned} \tag{7.4}$$

The operator L is a symmetric hyperbolic operator with constant coefficients and coefficient of ∂_t equal to I. It follows that L generates a unitary evolution on the spaces $H^s(\mathbb{R}^{d+1}_{x,\phi})$. Thus for all s and for all t smaller than the maximal time of existence,

$$\begin{aligned} \big\|\varepsilon^p \mathcal{E}^\delta(t)\big\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} \;\le\; &\big\|\varepsilon^p \mathcal{E}^\delta(0)\big\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} \\ &+ \int_0^t \Big\|L\Big(\partial_t, \partial_x + \frac{\xi_0}{\varepsilon}\partial_\phi\Big)\varepsilon^p \mathcal{E}^\delta(\sigma)\Big\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})}\, d\sigma. \end{aligned} \tag{7.5}$$

The key is to estimate the integral on the right-hand side of (7.5) using the expression (7.4).

7.1. Estimate for the R^δ term in (7.4). Estimate (6.3) implies that

$$\begin{aligned} \Big\|R^\delta\Big(\varepsilon, \varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big)\Big\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} &= \big\|\mathrm{R}^\delta(\varepsilon, \varepsilon t, x, z)\big\|_{H^s(\mathbb{R}^{d+1}_{x,z})} \\ &\le C\,\varepsilon^{p+1}\Big(\frac{\varepsilon}{\delta^2} + \varepsilon^p\Big). \end{aligned} \tag{7.6}$$

7.2. Estimate for the $(I - \chi^\delta(D_z))\,\Phi(\varepsilon^p U^\delta(\varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi))$ term in (7.4). As in the derivation of (7.6), the translation invariance of the H^s norm yields

$$\begin{aligned} &\Big\|\big(I - \chi^\delta(D_z)\big)\Phi\Big(\varepsilon^p U^\delta\Big(\varepsilon t, t, x, \frac{t\tau_0}{\varepsilon} + \phi\Big)\Big)\Big\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} \\ &\qquad = \big\|\big(I - \chi^\delta(D_z)\big)\Phi(\varepsilon^p \mathrm{U}^\delta(\varepsilon t, x, z))\big\|_{H^s(\mathbb{R}^{d+1}_{x,z})}. \end{aligned}$$

Express $\Phi(\varepsilon^p \mathrm{U}^\delta)$ as the sum of two terms,

$$\Phi(\varepsilon^p \mathrm{U}^\delta) = \Phi_J(\varepsilon^p \mathrm{U}^\delta) + \big(\Phi - \Phi_J\big)(\varepsilon^p \mathrm{U}^\delta) = \varepsilon^{p+1}\Phi_J(\mathrm{U}^\delta) + \varepsilon^{2p+1}\Xi(\varepsilon^p, \mathrm{U}^\delta), \tag{7.7}$$

where we have used (6.9) to derive the second equality.

Since $\Xi(\varepsilon^p, \mathrm{U}^\delta(\varepsilon t, x, z)$ is uniformly bounded in $H^s(\mathbb{R}^{d+1}_{x,z})$ it follows that

$$\big\|(I-\chi^\delta(D_z))\varepsilon^{2p+1}\Xi(\varepsilon^p,\mathrm{U}^\delta)\big\|_{H^s(\mathbb{R}^{d+1}_{x,z})} \;\le\; C\,\varepsilon^{2p+1}. \tag{7.8}$$

Since $\Phi_J(\mathrm{U}^\delta(\varepsilon t, t, x, \frac{t\tau}{\varepsilon}+\phi))$ are uniformly bounded in $\mathbb{B}^{m,\infty}$, Lemma 5.2 implies that

$$\big\|(I-\chi^\delta(D_z))\varepsilon^{p+1}\Phi_J(U^\delta)\big\|_{H^s(\mathbb{R}^{d+1}_{x,z})} \le C\,\sqrt{\delta}\,\varepsilon^{p+1}. \tag{7.9}$$

Adding (7.8) and (7.9) shows that

$$\Big\|\big(I-\chi^\delta(D_z)\big)\,\Phi\Big(\varepsilon^p U^\delta\Big(\varepsilon t, t, x, \frac{t\tau_0}{\varepsilon}+\phi\Big)\Big)\Big\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} \le C\,\varepsilon^{p+1}(\sqrt{\delta}+\varepsilon^p). \tag{7.10}$$

7.3. Estimate for the $\Phi(\varepsilon^p U^\delta)-\Phi(\varepsilon^p\mathcal{V})$ term in (7.4). Using (6.8) yields

$$\Phi(\varepsilon^p U^\delta)-\Phi(\varepsilon^p\mathcal{V}) = \varepsilon^{p+1}\big[\Psi(\varepsilon^p,U^\delta)-\Psi(\varepsilon^p,\mathcal{V})\big] = \varepsilon^{p+1}\big[\Psi(\varepsilon^p,U^\delta)-\Psi(\varepsilon^p,U^\delta+\mathcal{E}^\delta)\big].$$

The proof of Proposition 5.1 shows that

$$\sup_{0\le t\le T/\varepsilon,\ 0<\varepsilon<\delta\le\delta_0}\ \Big\|U^\delta\Big(\varepsilon,\varepsilon t,t,x,\frac{t\tau_0}{\varepsilon}+\phi\Big)\Big\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})}<\infty.$$

Schauder's lemma then implies that *as long as* $\|\mathcal{E}^\delta\|_{H^s}\le 1$,

$$\varepsilon^{p+1}\big\|\Psi(\varepsilon^p,U^\delta+\mathcal{E}^\delta)-\Psi(\varepsilon^p,U^\delta)\big\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} \le C\,\varepsilon^{p+1}\,\|\mathcal{E}^\delta(t)\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})}. \tag{7.11}$$

Combining estimates (7.6), (7.10), and (7.11) yields the following estimate, valid *as long as* $\|\mathcal{E}^\delta\|_{H^s}\le 1$:

$$\Big\|L\Big(\partial_t,\partial_x+\frac{\xi_0}{\varepsilon}\partial_\phi\Big)\,\varepsilon^p\mathcal{E}^\delta\Big\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} \;\le\; C\,\varepsilon^{p+1}\,\Big(\frac{\varepsilon}{\delta^2}+\varepsilon^p+\sqrt{\delta}+\|\mathcal{E}^\delta(t)\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})}\Big). \tag{7.12}$$

7.4. Estimate for $\mathcal{E}^\delta(0)$. The initial value of $\mathcal{E}^\delta(0)$ comes from the correctors in U^δ,

$$\mathcal{E}^\delta(0) = f(x,\phi)-\mathrm{U}^\delta(0,x,\phi) = \varepsilon\mathrm{U}_1^\delta(0,x,\phi)+\varepsilon^2\mathrm{U}_2^\delta(0,x,\phi).$$

Using (6.4) and the fact that $\varepsilon\le\delta$ yields

$$\|\mathcal{E}^\delta(0)\|_{H^s(\mathbb{R}^{d+1})} \;\le\; C\Big(\frac{\varepsilon}{\delta}+\frac{\varepsilon^2}{\delta^2}\Big) \;\le\; \frac{C\varepsilon}{\delta}. \tag{7.13}$$

8. End of proof. We now use the previous results to bound the error $\mathcal{E}^\delta$ between the exact solution $\mathcal{V}$ and the approximate solution defined in terms of U^δ.

Fix $s>(d+1)/2$. Then as long as $\|\mathcal{E}^\delta\|_{H^s}\le 1$, inserting (7.12) and (7.13) into (7.5) yields

$$\big\|\,\varepsilon^p\,\mathcal{E}^\delta(t)\,\big\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} \;\le\; \frac{C\varepsilon^{p+1}}{\delta}+C\int_0^t\varepsilon^{p+1}\,\Big(\frac{\varepsilon}{\delta^2}+\varepsilon^p+\sqrt{\delta}+\|\mathcal{E}^\delta(\sigma)\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})}\Big)\,d\sigma.$$

Estimate the integral of the constant term using $t \le \underline{T}/\varepsilon$ to find

$$\int_0^t \varepsilon^{p+1}\left(\frac{\varepsilon}{\delta^2}+\varepsilon^p+\sqrt{\delta}\right)\,d\sigma \le \frac{\underline{T}\,\varepsilon^{p+1}}{\varepsilon}\left(\frac{\varepsilon}{\delta^2}+\varepsilon^p+\sqrt{\delta}\right) \;\le\; C\,\varepsilon^p\Big(\frac{\varepsilon}{\delta^2}+\varepsilon^p+\sqrt{\delta}\Big).$$

Combine the last two estimates using $\frac{\varepsilon}{\delta^2} \ge \frac{\varepsilon}{\delta}$ and divide by ε^p to find

$$\big\|\,\mathcal{E}^\delta(t)\,\big\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} \le C\left(\frac{\varepsilon}{\delta^2}+\varepsilon^p+\sqrt{\delta}\right)+C\,\varepsilon\int_0^t \|\mathcal{E}^\delta(\sigma)\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})}\,d\sigma. \tag{8.1}$$

We need to show that $\mathcal{E}^\delta$ exists for $0 \le T \le \underline{T}/\varepsilon$ and that $\sup_{t\in[0,\underline{T}/\varepsilon]} \|\mathcal{E}^\delta(t)\|_{H^s}$ converges to zero as $\delta \to 0$. The integral inequality (8.1) leads to both of these goals by the "as long as" argument.

For any $0 < \varepsilon < \delta < \delta_0$, define $T_* = T_*(\varepsilon, s, \delta)$ by $T_* = \underline{T}/\varepsilon$ if $U^\delta(\varepsilon, T, t, x, z)$ exists for $0 \le t \le \underline{T}/\varepsilon$ and $\sup_{0\le t\le \underline{T}/\varepsilon} \|\mathcal{E}^\delta(t)\|_{H^s} < 1$. Otherwise define

$$T_* \;:=\; \inf\big\{t \,:\, \|\mathcal{E}^\delta(t)\|_{H^s} = 1\big\}.$$

Gronwall's inequality applied to (8.1) implies that for $0 \le t \le T_*$,

$$\|\mathcal{E}^\delta(t)\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} \le C(s)\left(\frac{\varepsilon}{\delta^2}+\varepsilon^p+\sqrt{\delta}\right)e^{C(s)\varepsilon t} \le C(s)\left(\frac{\varepsilon}{\delta^2}+\varepsilon^p+\sqrt{\delta}\right). \tag{8.2}$$

Now we can chose δ as a function of ε. Balancing the ε/δ^2 and $\sqrt{\delta}$ terms in (8.2) yields

$$\delta = \varepsilon^{0.4} \qquad \text{and} \qquad \frac{\varepsilon}{\delta^2} = \sqrt{\delta} = \varepsilon^{1/5}.$$

Then (8.2) yields for $\delta = \varepsilon^{0.4}$

$$\|\mathcal{E}^\delta(t)\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} \;\le\; C(s)\,\varepsilon^{\min\{p\,,\,1/5\}}. \tag{8.3}$$

Choose $\varepsilon(s) > 0$ so that

$$C(s)\,\varepsilon(s)^{\min\{p\,,\,1/5\}} \;<\; \frac{1}{2}. \tag{8.4}$$

Combining (8.3) and (8.4) shows that for $0 \le t \le T_*$, $0 < \varepsilon < \varepsilon(s)$, and $\delta = \varepsilon^{0.4}$

$$\|\mathcal{E}^\delta(t)\|_{H^s(\mathbb{R}^{d+1}_{x,\phi})} < \frac{1}{2}. \tag{8.5}$$

If $T_* < \underline{T}$, setting $t = T_*$ violates the definition of T_*. It follows that for $0 < \varepsilon < \varepsilon(s)$, $\mathcal{E}^{e^{0.4}}(t)$ has H^s norm less than $1/2$ for $0 \le t \le \underline{T}/\varepsilon$.

This proves the solvability for $0 \le t \le \underline{T}/\varepsilon$ of the initial value problem defining $\mathcal{V} \in C([0,T]\,;\,H^s)$. That the solution belongs to $C([0,\underline{T}/\varepsilon]\,;\,\cap_s H^s)$ is then a consequence of standard semilinear hyperbolic theory and the regularity of the initial data for $\mathcal{V}$.

In addition, since (8.5) holds, the "as long as" argument works, and it follows that inequality (8.3) is valid for $0 \le t \le \underline{T}/\varepsilon$, provided that $\varepsilon < \varepsilon(s)$. Since s is arbitrary this proves the convergence asserted in Proposition 7.1. ∎

8.1. The main theorems. Combining Propositions 4.3, 5.1, and 7.1 and Corollary 4.12 proves the following result.

THEOREM 8.1 (main theorem). *Assume that the initial data in* (2.1) *satisfy Assumptions* 1.5 *and* 1.8. *Let* $\mathrm{U}_0 = \pi(\beta)\mathrm{U}_0 \in C([0,T_*[\,;\mathbb{A}(\mathbb{R}^{1+d})$ *be the maximal solution of the principal profile equation* (4.2) *with initial value* f. *Let* $\mathcal{V}(\varepsilon,t,x,\phi) \in \cap_s C([0,T'_*[\,;H^s(\mathbb{R}^{d+1}))$ *denote the maximal solution of the initial value problem* (2.10), (2.11) *defining the exact profile.*

Then, for any $\underline{T} < T_*$ *there is an* $\varepsilon(\underline{T}) > 0$ *so that for* $0 < \varepsilon < \varepsilon(\underline{T})$ *the solution* u^ε *of the initial value problem* (2.1) *exists for* $0 \le t \le \underline{T}/\varepsilon$, $T'_* \ge \underline{T}/\varepsilon$, *and* u^ε *is given by*

$$u^\varepsilon \;=\; \varepsilon^p \mathcal{V}\Big(\varepsilon,t,x,\frac{x.\xi_0}{\varepsilon}\Big).$$

In addition the asymptotic behavior as $\varepsilon \to 0$ *is given by*

$$u^\varepsilon \;\sim\; \varepsilon^p \mathrm{U}_0\Big(\varepsilon,\varepsilon t,x-\mathbf{v}t,\frac{t\tau_0+x.\xi_0}{\varepsilon}\Big)$$

in the sense that for all s, *as* $\varepsilon \to 0$

$$\sup_{0\le t\le \underline{T}/\varepsilon}\Big\|\mathcal{V}(\varepsilon,t,x,\phi)-\mathrm{U}_0\Big(\varepsilon t,x-\mathbf{v}t,\frac{t\tau_0}{\varepsilon}+\phi\Big)\Big\|_{H^s(\mathbb{R}^{1+d}_{x,\phi})} \;\le C(s)\,\varepsilon^{\min\{p,1/5\}}. \tag{8.6}$$

Proof of Theorem 1.10. Theorem 1.10 is an immediate consequence of this result. Simply write

$$u^\varepsilon - u^\varepsilon_{\mathrm{approx}} \;=\; \varepsilon^p \mathcal{V}\Big(\varepsilon,t,x,\frac{x.\xi_0}{\varepsilon}\Big) - \varepsilon^p \mathrm{U}_0\Big(\varepsilon t,x-\mathbf{v}t,\frac{t\tau_0+x.\xi_0}{\varepsilon}\Big).$$

Then the estimate (1.12) follows from (8.6). ∎

Note that the constant field V_d has a nonzero ∂_t component. It acts differently on the two terms in the expression for the error. That is why we have a reduced set of derivatives in the error estimate of Theorem 1.10.

REFERENCES

[1] D. ALTERMAN, *Diffractive Nonlinear Geometric Optics for Short Pulses*, Ph.D. thesis, Department of Mathematics, University of Michigan, Ann Arbor, MI, 1999.

[2] D. ALTERMAN AND J. RAUCH, *Diffractive short pulse asymptotics for nonlinear wave equations*, Phys. Lett. A., 264 (2000), pp. 390–395.

[3] D. ALTERMAN AND J. RAUCH, *Correcting the failure of the slowly varying amplitude approximation for short pulses*, in Ultrafast Phenomena XII, T. Elsaesser, S. Mukamel, M. M. Murnane, and N. F. Scherer, eds., Springer-Verlag, New York, 2001, pp. 71–73.

[4] D. ALTERMAN AND J. RAUCH, *The linear diffractive pulse equation*, Methods Appl. Anal., 7 (2000), pp. 263–274.

[5] D. ALTERMAN AND J. RAUCH, *Nonlinear geometric optics for short pulses*, J. Differential Equations, 178 (2002), pp. 437–465.

[6] K. BARRAILH AND D. LANNES, *A general framework for diffractive optics and its applications to lasers with large spectrums and short pulses*, SIAM J. Math. Anal., 34 (2003), pp. 636–674.

[7] R. CARLES AND J. RAUCH, *Focusing of spherical nonlinear pulses in* $\mathbb{R}^{1+3}$, Proc. Amer. Math. Soc., 130 (2002), pp. 791–804.

[8] R. CARLES AND J. RAUCH, *Absorption d'impulsions non-linéaires radiales focalisantes dans* $\mathbb{R}^{1+3}$, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 985–990.

[9] R. CARLES AND J. RAUCH, *Diffusion d'impulsions non-linéaires radiales focalisantes dans* $\mathbb{R}^{1+3}$, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 1077–1082.

[10] T. Brabec and F. Krausz, *Nonlinear optical pulse propagation in the single-cycle regime*, Phys. Rev. Lett., 78 (1997), pp. 3282–3285.
[11] P. Donnat, J.-L. Joly, G. Métivier, and J. Rauch, *Diffractive Nonlinear Geometric Optics*, in Séminaire sur les Equations aux Dérivées Partielles (1995–1996), Exp. 17, École Polytéchnique, Paris, 1996.
[12] E. Dumas, *Periodic multiphase diffractive optics with curved phases*, Indiana Math. J., to appear.
[13] E. Esarey, P. Sprangle, M. Pilloff, and J. Krall, *Theory and group velocity of ultrashort, tightly focused laser pulses*, J. Opt. Soc. Amer. B Opt. Phys., 12 (1995), pp. 1695–1703.
[14] S. Feng, H. G. Winful, and R. W. Hellwarth, *Spatiotemporal evolution of focused single-cycle electromagnetic pulses*, Phys. Rev. E, 59 (1999), pp. 4630–4649.
[15] R. W. Hellwarth and P. Nouchi, *Focused one-cycle electromagnetic pulses*, Phys. Rev. E, 54 (1996), pp. 889–895.
[16] C. Hile, *Comparisons between Maxwell's equations and an extended nonlinear Schrödinger equation*, Wave Motion, 24 (1996), pp. 1–12.
[17] C. V. Hile and W. L. Kath, *Numerical solutions of Maxwell's equations for nonlinear-optical pulse propagation*, J. Opt. Soc. Amer. B Opt. Phys., 13 (1996), pp. 1135–1145.
[18] J.-L. Joly, G. Métivier, and J. Rauch, *Coherent and focusing multidimensional nonlinear geometric optics*, Ann. Sci. École Norm. Sup. (4), 28 (1995), pp. 51–113.
[19] J.-L. Joly, G. Metivier, and J. Rauch, *Diffractive nonlinear geometric optics with rectification*, Indiana Univ. Math. J., 47 (1998), pp. 1167–1242.
[20] A. E. Kaplan, *Diffraction-induced transformation of near-cycle and subcycle pulses*, J. Opt. Soc. Amer. B Opt. Phys., 15 (1998), pp. 951–956.
[21] M. A. Porras, *Ultrashort pulsed Gaussian light beams*, Phys. Rev. E, 58 (1998), pp. 1086–1093.
[22] J. K. Ranka and A. L. Gaeta, *Breakdown of the slowly varying envelope approximation in the self-focusing of ultrashort pulses*, Opt. Lett., 23 (1998), pp. 534–536.
[23] J. Rauch, *An L^2 proof that H^s is invariant under nonlinear maps for $s > n/2$*, in Global Analysis: Analysis on Manifolds, T. M. Rassias, ed., Teubner, Leipzig, 1983, pp. 301–305.
[24] J. Rauch, *Lectures on geometric optics*, in Hyperbolic Equations and Frequency Interactions, L. Caffarelli and W. E, eds., IAS/Park City Math. Ser. 5, AMS, Providence, RI, 1999, pp. 383–466.
[25] J. E. Rothenberg, *Space-time focusing: Breakdown of the slowly varying envelope approximation in the self-focusing of femtosecond pulses*, Opt. Lett., 17 (1992), pp. 1340–1342.
[26] A. Yoshikawa, *Solutions containing a large parameter of a quasi-linear hyperbolic system of equations and their nonlinear geometric optics approximation*, Trans. Amer. Math. Soc., 340 (1993), pp. 103–126.
[27] A. Yoshikawa, *Asymptotic expansions of the solutions to a class of quasilinear hyperbolic initial value problems*, J. Math. Soc. Japan, 47 (1995), pp. 227–252.
[28] R. W. Ziolkowski, *Localized transmission of electromagnetic energy*, Phys. Rev. A, 39 (1989), pp. 2005–2033.